

# DILATIONS, LINEAR MATRIX INEQUALITIES, THE MATRIX CUBE PROBLEM AND BETA DISTRIBUTIONS

J. WILLIAM HELTON<sup>1</sup>, IGOR KLEP<sup>2</sup>, SCOTT MCCULLOUGH<sup>3</sup>, AND MARKUS SCHWEIGHOFER

**ABSTRACT.** An operator  $C$  on a Hilbert space  $\mathcal{H}$  dilates to an operator  $T$  on a Hilbert space  $\mathcal{K}$  if there is an isometry  $V : \mathcal{H} \rightarrow \mathcal{K}$  such that  $C = V^*TV$ . A main result of this paper is, for a positive integer  $d$ , the simultaneous dilation, up to a sharp factor  $\vartheta(d)$ , expressed as a ratio of  $\Gamma$  functions for  $d$  even, of all  $d \times d$  symmetric matrices of operator norm at most one to a collection of *commuting* self-adjoint contraction operators on a Hilbert space.

Dilating to commuting operators has consequences for the theory of linear matrix inequalities (LMIs). Given a tuple  $A = (A_1, \dots, A_g)$  of  $\nu \times \nu$  symmetric matrices,  $L(x) := I - \sum A_j x_j$  is a *monic linear pencil* of size  $\nu$ . The solution set  $\mathcal{S}_L$  of the corresponding linear matrix inequality, consisting of those  $x \in \mathbb{R}^g$  for which  $L(x) \succeq 0$ , is a *spectrahedron*. The set  $\mathcal{D}_L$  of tuples  $X = (X_1, \dots, X_g)$  of symmetric matrices (of the same size) for which  $L(X) := I - \sum A_j \otimes X_j$  is positive semidefinite, is a *free spectrahedron*. It is shown that any tuple  $X$  of  $d \times d$  symmetric matrices in a bounded free spectrahedron  $\mathcal{D}_L$  dilates, up to a scale factor depending only on  $d$ , to a tuple  $T$  of *commuting* self-adjoint operators with joint spectrum in the corresponding spectrahedron  $\mathcal{S}_L$ . From another viewpoint, the scale factor measures the extent that a positive map can fail to be completely positive.

Given another monic linear pencil  $\tilde{L}$ , the inclusion  $\mathcal{D}_L \subseteq \mathcal{D}_{\tilde{L}}$  obviously implies the inclusion  $\mathcal{S}_L \subseteq \mathcal{S}_{\tilde{L}}$  and thus can be thought of as its free relaxation. Determining if one free spectrahedron contains another can be done by solving an explicit LMI and is thus computationally tractable. The scale factor for commutative dilation of  $\mathcal{D}_L$  gives a precise measure of the worst case error inherent in the free relaxation, over all monic linear pencils  $\tilde{L}$  of size  $d$ .

The set  $\mathfrak{C}^{(g)}$  of  $g$ -tuples of symmetric matrices of norm at most one is an example of a free spectrahedron known as the free cube and its associated spectrahedron is the cube  $[-1, 1]^g$ . The free relaxation of the the NP-hard inclusion problem  $[-1, 1]^g \subseteq \mathcal{S}_L$  was introduced by Ben-Tal and Nemirovski. They obtained the lower bound  $\vartheta(d)$ , expressed as the solution of an optimization problem over diagonal matrices of trace norm 1, for the divergence between the original and relaxed problem. The result here on simultaneous dilations of contractions proves this bound is sharp. Determining an analytic formula for  $\vartheta(d)$  produces, as a by-product, new probabilistic results for the binomial and beta distributions.

---

*Date:* May 10, 2016.

*2010 Mathematics Subject Classification.* 47A20, 46L07, 13J30 (Primary); 60E05, 33B15, 90C22 (Secondary).

*Key words and phrases.* dilation, completely positive map, linear matrix inequality, spectrahedron, free spectrahedron, matrix cube problem, binomial distribution, beta distribution, robust stability, free analysis.

<sup>1</sup>Research supported by the NSF grant DMS 1201498, and the Ford Motor Co.

<sup>2</sup>Supported by the Marsden Fund Council of the Royal Society of New Zealand. Partially supported by the Slovenian Research Agency grants P1-0222 and L1-6722.

<sup>3</sup>Research supported by the NSF grant DMS-1361501.

## 1. INTRODUCTION

Free analysis [KVV14] and the theory of free functions and free sets traces its roots back to the work of Taylor [Tay72, Tay73]. Free functions generalize the notion of polynomials in  $g$ -tuples of freely noncommuting variables and free sets are appropriately structured subsets of the union, over  $d$ , of  $g$ -tuples of  $d \times d$  matrices. The current interest in the subject arises in large part due to its applications in free probability [Voi04, Voi10], systems engineering and connections to optimization [Bal11, BGR90, BGFB94, dOHMP09, SIG96] and operator algebras and systems [Arv69, Arv72, Pau02, BLM04, Pis03, Dav12, DK+]. The main branch of convex optimization to emerge in the last 20 years, semidefinite programming [Nem06], is based on linear pencils, linear matrix inequalities (LMIs) and spectrahedra [BGFB94, WSV00]. The book [BPR13] gives an overview of the substantial theory of LMIs and spectrahedra and the connection to real algebraic geometry. A linear pencil  $L$  is a simple special case of a free function and is of special interest because the free spectrahedron  $\mathcal{D}_L = \{X : L(X) \succeq 0\}$  is evidently convex and conversely an algebraically defined free convex set is a free spectrahedron [EW97, HM12]. In this article the relation between inclusions of spectrahedra and inclusions of the corresponding free spectrahedra is explored using operator theoretic ideas. The analysis leads to new dilation theoretic results and to new probabilistic results and conjectures which can be read independently of the rest of this article by skipping now to Section 1.8. It also furnishes a complete solution to the matrix cube problem of Ben-Tal and Nemirovski [B-TN02], which contains, as a special case, the  $\frac{\pi}{2}$ -Theorem of Nesterov [Nes97] and which in turn is related to the symmetric Grothendieck Inequality.

A central topic of this paper is dilation, up to a scale factor, of a tuple  $X$  of  $d \times d$  symmetric matrices in a free spectrahedron  $\mathcal{D}_L$  to tuples  $T$  of *commuting* self-adjoint operators with joint spectrum in the corresponding spectrahedron  $\mathcal{S}_L$ . We shall prove that these scaled commutative dilations exist and the scale factor describes the error in the free relaxation  $\mathcal{D}_L \subseteq \mathcal{D}_{\tilde{L}}$  of the spectrahedral inclusion problem  $\mathcal{S}_L \subseteq \mathcal{S}_{\tilde{L}}$ . The precise results are stated in Section 1.3. As a cultural note these scale factors can be interpreted as the amount of modification required to make a positive map completely positive; see Section 1.4.

In this paper we completely analyze the free cubes,  $\mathfrak{C}^{(g)}$ , the free spectrahedra consisting of  $g$ -tuples of symmetric matrices of norm at most one; the corresponding spectrahedron is the cube  $[-1, 1]^g$ . We show that, for each  $d$ , there exists a collection  $\mathcal{C}_d$  of commuting self-adjoint contraction operators on a Hilbert space, such that, up to the scale factor  $\vartheta(d)$ , any  $d \times d$  symmetric contraction matrix dilates to  $T$  in  $\mathcal{C}_d$ ; see Section 1.1. Moreover, we give a formula for the optimal scale factor  $\vartheta(d)$ ; see Section 1.2. As a consequence, we recover the error bound given by Ben-Tal and Nemirovski for the computationally tractable free relaxation of the NP-hard cube inclusion problem. Further, we show that this bound is best possible, see Section 1.5.

Proof of sharpness of the error bound  $\vartheta(d)$  and giving a formula for  $\vartheta(d)$  requires concatenating all of the areas we have discussed and it requires all but a few sections of this paper. For example, finding a formula for  $\vartheta(d)$  required new results for the binomial and beta

distributions and necessitated a generalization of Simmons' Theorem [Sim1894] (cf. [PR07]) to Beta distributions. Our results and conjectures in probability-statistics appear in Section 1.8.

The rest of the introduction gives detailed statements of the results just described and a guide to their proofs.

**1.1. Simultaneous dilations.** Denote by  $\mathbb{N} := \{1, 2, 3, \dots\}$  the set of positive integers and by  $\mathbb{R}$  the set of real numbers. For  $n \in \mathbb{N}$ , denote by  $\mathbb{S}_n$  the set of symmetric  $n \times n$  matrices with entries from  $\mathbb{R}$ . A matrix  $X \in \mathbb{S}_n$  **dilates** to an operator  $T$  on a Hilbert space  $\mathcal{H}$  if there is an isometry  $V : \mathbb{R}^n \rightarrow \mathcal{H}$  such that  $X = V^*TV$ . Alternately, one says that  $X$  is a **compression** of  $T$ . A tuple  $X \in \mathbb{S}_n^g$  **dilates** to a tuple  $T = (T_1, \dots, T_g)$  of bounded operators on a Hilbert space  $\mathcal{H}$  if there is an isometry  $V : \mathbb{R}^n \rightarrow \mathcal{H}$  such that  $X = V^*TV$  (in the sense that  $X_j = V^*T_jV$ ). In other words,  $T$  has the form

$$T_i = \begin{pmatrix} X_i & *_{i-1} \\ *_{i-1} & *_{i-2} \end{pmatrix}$$

where the  $*_i$  are bounded operators between appropriate Hilbert spaces.

One of the oldest dilation theorems is due to Naimark [Nai43]. In its simplest form it dilates a tuple of (symmetric) positive semidefinite matrices (of the same size) which sum to the identity to a tuple of commuting (symmetric) projections which sum to the identity. It has modern applications in the theory of frames from signal processing. In this direction, perhaps the most general version of the Naimark Dilation Theorem dilates a (possibly nonselfadjoint) operator valued measure to a (commuting) projection valued measure on a Banach space. The most general and complete reference for this result and its antecedents is [HLLL14a] with [HLLL14b] being an exposition of the theory. See also [LS13]. A highly celebrated dilation result in operator theory is the power dilation theorem of Sz.-Nagy [SzN53] which, given a contraction  $X$ , constructs a unitary  $U$  such that  $X^n$  dilates to  $U^n$  for natural numbers  $n$ . That von Neumann's inequality is an immediate consequence gives some measure of the power of this result. The two variable generalization of the Sz.-Nagy dilation theorem, the power dilation of a commuting pair of contractions to a commuting pair of unitaries, is known as the commutant lifting theorem (there are counterexamples to commutant lifting for more than two contractions) and is due to Ando, Foias-Sz.-Nagy, Sarason. It has major applications to linear systems engineering; see [Bal11, FFGK98, BGR90] as samples of the large literature. The (latest revision of the) classic book [SzNFBK10, Chapter 1.12] contains further remarks on the history of dilations. Power dilations up to a scale factor  $K$  are often called  $K$ -spectral dilations and these are a highly active area of research. An excellent survey article is [BB13].

The connections between dilations and completely positive maps were exposed most famously in the work of Arveson [Arv69, Arv72]. Presently, dilations and completely positive maps appear in many contexts. For examples, they are fundamental objects in the theory of operator algebras, systems and spaces [Pau02] as well as in quantum computing and quantum information theory [NC11]. In the articles [HKM12, HKM13], the theory of completely positive maps was used to systematically study free relaxations of spectrahedral inclusion problems

which arise in semidefinite programming [Nem06, WSV00] and systems engineering [B-TN02] for instance. In this article, dilation theory is used to measure the degree to which a positive map can fail to be completely positive, equivalently the error inherent in free relaxations of spectrahedral inclusion.

The dilation constant  $\vartheta(d)$  for dilating  $d \times d$  contractions to commuting contractions operators is given by the following optimization problem [B-TN02] which is potentially of independent interest.

$$(1.1) \quad \frac{1}{\vartheta(d)} := \min_{\substack{a \in \mathbb{R}^d \\ |a_1| + \dots + |a_d| = d}} \int_{S^{d-1}} \left| \sum_{i=1}^d a_i \xi_i^2 \right| d\xi = \min_{\substack{B \in \mathbb{S}_d \\ \text{tr } |B| = d}} \int_{S^{d-1}} |\xi^* B \xi| d\xi$$

where the unit sphere  $S^{d-1} \subseteq \mathbb{R}^d$  (having dimension  $d-1$ ) is equipped with the uniform probability measure (i.e., the unique rotation invariant measure of total mass 1). That  $\vartheta(d)$  can be expressed using incomplete beta functions will be seen in Section 1.2. Evidently  $\vartheta(d) \geq 1$ .

A self-adjoint operator  $Y$  on  $\mathcal{H}$  is a **contraction** if  $I \pm Y \succeq 0$  or equivalently  $\|Y\| \leq 1$ .

**Theorem 1.1** (Simultaneous Dilation). *Let  $d \in \mathbb{N}$ . There is a Hilbert space  $\mathcal{H}$ , a family  $\mathcal{C}_d$  of commuting self-adjoint contractions on  $\mathcal{H}$ , and an isometry  $V : \mathbb{R}^d \rightarrow \mathcal{H}$  such that for each symmetric  $d \times d$  contraction matrix  $X$  there exists a  $T \in \mathcal{C}_d$  such that*

$$\frac{1}{\vartheta(d)} X = V^* T V.$$

*Moreover,  $\vartheta(d)$  is the smallest such constant in the sense that if  $\vartheta' \in \mathbb{R}$  satisfies  $1 \leq \vartheta' < \vartheta(d)$ , then there is  $g \in \mathbb{N}$  and a  $g$ -tuple of  $d \times d$  symmetric contractions  $X$  such that  $\frac{1}{\vartheta'} X$  does not dilate to a  $g$ -tuple of commuting self-adjoint contractions on a Hilbert space.*

*Proof.* The first part of Theorem 1.1 is stated and proved as Theorem 5.9. The optimality of  $\vartheta(d)$  is proved as part of Theorem 5.10. The Hilbert space  $\mathcal{H}$ , isometry  $V$  and collection  $\mathcal{C}_d$  are all explicitly constructed. See equations (3.2), (3.3) and (3.4). ■

**1.2. Solution of the minimization problem (1.1).** In this section matrices  $B$  which produce the optimum in Equation (1.1) are described and a formula for  $\vartheta(d)$  is given in terms of Beta functions. Recall, the incomplete beta function is, for real arguments  $\alpha, \beta > 0$ , and an additional argument  $p \in [0, 1]$ , defined by

$$B_p(\alpha, \beta) = \int_0^p x^{\alpha-1} (1-x)^{\beta-1} dx.$$

The Euler beta function is  $B(\alpha, \beta) = B_1(\alpha, \beta)$  and the regularized (incomplete) beta function is

$$I_p(\alpha, \beta) = \frac{B_p(\alpha, \beta)}{B(\alpha, \beta)} \in [0, 1].$$

The minimizing matrices  $B$  to (1.1) will have only two different eigenvalues. For nonnegative numbers  $a, b$  and  $s, t \in \mathbb{N}$ , let  $J(s, t; a, b) = aI_s \oplus (-b)I_t$  denote the  $d \times d$  diagonal matrix  $J(s, t; a, b)$  with first  $s$  diagonal entries  $a$  and last  $t$  diagonal entries  $-b$ .

The description of the solution to (1.1) depends on the parity of  $d$ .

**Theorem 1.2.** *If  $d$  is an even positive integer, then*

$$(1.2) \quad \frac{1}{\vartheta(d)} = \int_{S^{d-1}} \left| \xi^* J\left(\frac{d}{2}, \frac{d}{2}; 1, 1\right) \xi \right| d\xi = 2I_{\frac{1}{2}}\left(\frac{d}{4}, \frac{d}{4} + 1\right) - 1 = \frac{\Gamma\left(\frac{1}{2} + \frac{d}{4}\right)}{\sqrt{\pi} \Gamma\left(1 + \frac{d}{4}\right)}$$

where  $\Gamma$  denotes the Euler gamma function. In particular, the minimum in (1.1) occurs at a  $B = J(s, t; a, b)$  with  $s = t = \frac{d}{2}$  and  $a = b = 1$ .

In the case that  $d$  is odd, there exist  $a, b \geq 0$  such that

$$(1.3) \quad \frac{1}{\vartheta(d)} = \int_{S^{d-1}} \left| \xi^* J\left(\frac{d+1}{2}, \frac{d-1}{2}; a, b\right) \xi \right| d\xi$$

$$(1.4) \quad = 2I_{\frac{a}{a+b}}\left(\frac{d-1}{4}, \frac{d+1}{4} + 1\right) - 1 = 2I_{\frac{b}{a+b}}\left(\frac{d+1}{4}, \frac{d-1}{4} + 1\right) - 1,$$

and  $a\frac{d+1}{2} + b\frac{d-1}{2} = d$ . This last equation together with (1.4) uniquely determines  $a, b$ . Furthermore, the minimum in (1.1) occurs at a  $B = J(s, t; a, b)$  with  $s = \frac{d+1}{2}$  and  $t = \frac{d-1}{2}$ .

**1.2.1. Proof of Theorem 1.2.** The proof is involved but we now describe some of the ideas. A key observation is that the minimum defining  $\vartheta(d)$  in (1.1) can be taken over matrices of the form  $J(s, t; a, b)$ , instead of over all symmetric matrices  $B$  with  $\text{tr}(|B|) = d$ ; see Proposition 4.2. In addition,  $s + t = d$  and we may take  $as + bt = d = \text{tr}|J(s, t, a, b)|$ . The key identity connecting Beta functions to  $J$  is

$$(1.5) \quad \int_{S^{d-1}} |\xi^* J(s, t; a, b) \xi| d\xi = \int_{S^{d-1}} \left| a \sum_{j=1}^s \xi_j^2 - b \sum_{j=s+1}^d \xi_j^2 \right| d\xi \\ = \frac{2}{d} \left( as I_{\frac{a}{a+b}}\left(\frac{t}{2}, \frac{s}{2} + 1\right) + bt I_{\frac{b}{a+b}}\left(\frac{s}{2}, \frac{t}{2} + 1\right) \right) - 1,$$

which is verified in Section 6.

The optimality conditions for the optimization problem (1.1) (with  $J(s, t; a, b)$  replacing  $B \in \mathbb{S}_d$ ) are presented in Section 9, and the proof of Theorem 1.2 concludes in Section 12. ■

Bounds on the integral (1.4) representing  $\vartheta(d)$  when  $d$  is odd can be found below in Theorem 13.1.

**1.2.2. Coin flipping and Simmons' Theorem.** Theorem 1.2 is closely related to coin flipping. For example, when  $d$  is divisible by 4, the right hand side of (1.2) just becomes the probability of getting exactly  $\frac{d}{4}$  heads when tossing a fair coin  $\frac{d}{2}$  times, i.e.,

$$\binom{\frac{d}{2}}{\frac{d}{4}} \left(\frac{1}{2}\right)^{\frac{d}{2}}.$$

Furthermore, a core ingredient in analyzing the extrema of (1.5) or (1.1) as needed for  $\vartheta(d)$  is the following inequality.

**Theorem 1.3.** For  $d \in \mathbb{N}$  and  $s, t \in \mathbb{N}$  with  $s + t = d$ , if  $s \geq \frac{d}{2}$ , then

$$(1.6) \quad I_{\frac{s}{d}} \left( \frac{s}{2} + 1, \frac{t}{2} \right) \geq 1 - I_{\frac{s}{d}} \left( \frac{s}{2}, \frac{t}{2} + 1 \right).$$

*Proof.* See Section 10. ■

If  $s, d$  in Theorem 1.3 are both even, equivalently  $\frac{s}{2}$  and  $\frac{t}{2}$  are natural numbers, then (1.6) reduces to the following: toss a coin whose probability for head is  $\frac{s}{d} \geq \frac{1}{2}$ ,  $d$  times. Then the probability of getting fewer than  $s$  heads is *no more than* the probability of getting more than  $s$  heads. This result is known in classical probability as Simmon's Theorem [Sim1894].

Further probabilistic connections are described in Section 1.8.

**1.3. Linear matrix inequalities (LMIs), spectrahedra and general dilations.** In this section we discuss simultaneous dilation of tuples of symmetric matrices satisfying linear matrix inequalities to commuting self-adjoint operators.

For  $A, B \in \mathbb{S}_n$ , write  $A \preceq B$  (or  $B \succeq A$ ) to express that  $B - A$  is positive semidefinite (i.e., has only nonnegative eigenvalues). Given a  $g$ -tuple  $A = (A_1, \dots, A_g) \in \mathbb{S}_\nu^g$ , the expression

$$(1.7) \quad L_A(x) = I_\nu - \sum_{j=1}^g A_j x_j$$

is a **(monic) linear pencil** and  $L_A(x) \succeq 0$  is a **linear matrix inequality (LMI)**. Its solution set  $\mathcal{S}_{L_A} = \{x \in \mathbb{R}^g : L_A(x) \succeq 0\}$  is a **spectrahedron** (or an **LMI domain**) containing 0 in its interior [BPR13, BGFB94]. Conversely, each spectrahedron with non-empty interior can be written in this form after a translation [HV07]. Every polyhedron is a spectrahedron. For example, that the cube  $[-1, 1]^g$  in  $\mathbb{R}^g$  is an example of a spectrahedron, is seen as follows. Let  $E_j$  denote the  $g \times g$  diagonal matrix with a 1 in the  $(j, j)$  entry and zeros elsewhere and define  $C \in \mathbb{S}_{2g}^g$  by setting

$$(1.8) \quad C_j := \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \otimes E_j = \begin{pmatrix} E_j & 0 \\ 0 & -E_j \end{pmatrix}$$

for  $j \in \{1, \dots, g\}$ . The resulting spectrahedron  $\mathcal{S}_{L_C}$  is the cube  $[-1, 1]^g$ .

For  $n \in \mathbb{N}$  and tuples  $X \in \mathbb{S}_n^g$ , let

$$L_A(X) = I_{\nu n} - \sum_{j=1}^g A_j \otimes X_j, \quad \text{and} \\ \mathcal{D}_{L_A}(n) = \{X \in \mathbb{S}_n^g : L_A(X) \succeq 0\},$$

where  $\otimes$  is the Kronecker tensor product. The sequence  $\mathcal{D}_{L_A} = (\mathcal{D}_{L_A}(n))_n$  is a **free spectrahedron**. In particular,  $\mathcal{D}_{L_A}(1) = \mathcal{S}_{L_A}$  and  $\mathcal{D}_{L_C}(n)$  is the collection of  $g$ -tuples of  $n \times n$  symmetric contraction matrices. We call  $\mathfrak{C}^{(g)} := \mathcal{D}_{L_C}$  the **free cube** (in  $g$ -variables). Free spectrahedra are closely connected with operator systems for which [FP12, KPTT13, Arv08] are a few recent references.

1.3.1. *Dilations to commuting operators.* The general dilation problem is as follows: given a linear pencil  $L$  and a tuple  $X \in \mathcal{D}_L$ , does  $X$  dilate to a commuting tuple  $T$  of self-adjoint operators with joint spectrum in  $\mathcal{S}_L$ ?

Suppose  $L$  is a monic linear pencil in  $g$ -variables and the corresponding spectrahedron  $\mathcal{S}_L$  is bounded. Because there exist constants  $c$  and  $C$  such that

$$c\mathfrak{C}^{(g)} \subseteq \mathcal{D}_L \subseteq C\mathfrak{C}^{(g)},$$

from Theorem 1.1 it follows that for each  $n \in \mathbb{N}$ , and  $X \in \mathcal{D}_L(n)$  there exists a  $t \in \mathbb{R}_{>0}$ , a Hilbert space  $\mathcal{H}$ , a commuting tuple  $T$  of self-adjoint operators on  $\mathcal{H}$  with joint spectrum in  $\mathcal{S}_L$  and an isometry  $V : \mathbb{R}^n \rightarrow \mathcal{H}$  such that

$$X = V^* \frac{1}{t} T V.$$

The largest  $t$  such that for each  $X \in \mathcal{D}_L(n)$  the tuple  $tX$  dilates to a commuting tuple of self-adjoint operators with joint spectrum in  $\mathcal{S}_L$  is the **commutability index of  $L$** , denoted by  $\tau(L)(n)$ . (If  $\mathcal{S}_L$  is not bounded, then there need not be an upper bound on  $t$ .) The constants  $c, C$  and Theorem 1.6 below produce bounds on  $t$  depending only upon the monic pencil  $L$  and  $n$ .

1.3.2. *Spectrahedral inclusion problem.* Given two monic linear pencils  $L, \tilde{L}$  and corresponding spectrahedra, determine if the inclusion  $\mathcal{S}_L \subseteq \mathcal{S}_{\tilde{L}}$  holds. The article [HKM13] considered the free variable relaxation of this inclusion problem, dubbed the **free spectrahedral inclusion problem**: when does the containment  $\mathcal{D}_L \subseteq \mathcal{D}_{\tilde{L}}$  hold? In [HKM13, HKM12] it is shown, via an algorithm and using complete positivity, that any such free inclusion problem can be converted to an SDP feasibility problem (whose complexity status is unknown but is believed to be efficient to solve in theory and practice; cf. [WSV00, Ch. 8.4.4]). (See also Section 1.4 below.)

1.3.3. *Accuracy of the free relaxation.* Now that we have seen free spectrahedral inclusion problems are in principle solvable, what do they say about the original inclusion problem? Inclusion of free sets  $\mathcal{D}_L \subseteq \mathcal{D}_{\tilde{L}}$  implies trivially the inclusion of the corresponding classical spectrahedra  $\mathcal{S}_L \subseteq \mathcal{S}_{\tilde{L}}$ . Conversely, in the case that  $\mathcal{S}_{\tilde{L}}$  is bounded there exists  $c, C > 0$  such that

$$c\mathfrak{C}^{(g)} \subseteq \mathcal{D}_{\tilde{L}} \subseteq C\mathfrak{C}^{(g)},$$

and hence there exists an  $r \in \mathbb{R}_{>0}$  such that

$$\mathcal{S}_L \subseteq \mathcal{S}_{\tilde{L}} \quad \text{implies} \quad r\mathcal{D}_L \subseteq \mathcal{D}_{\tilde{L}}.$$

We call such an  $r$  an  $\mathcal{D}_L$ - $\mathcal{D}_{\tilde{L}}$ -**inclusion constant**. Theorem 1.6 and the constants  $c, C$  produce a lower bound on  $r$ . Let  $r(L, \tilde{L})$  denote the largest such  $r$  (if  $\mathcal{S}_{\tilde{L}}$  is not assumed bounded, then a largest  $r$  need not exist) and let

$$r(L)(d) := \min \{r(L, \tilde{L}) : \tilde{L} \text{ is of size } d \text{ and } \mathcal{S}_L \subseteq \mathcal{S}_{\tilde{L}}\}.$$

We call the sequence  $r(L) := (r(L)(d))_d$  the  $\mathcal{D}_L$ -**inclusion scale**.

The connection between spectrahedral inclusions and general dilations arises as follows:

**Theorem 1.4.** *Suppose  $L$  is a monic linear pencil and  $\mathcal{S}_L$  is bounded.*

- (1) *The commutability index for  $L$  equals its inclusion scale,  $\tau(L) = r(L)$ . That is  $\tau(L)(d)$  is the largest constant such that*

$$\tau(L)(d) \mathcal{D}_L \subseteq \mathcal{D}_{\tilde{L}}$$

*for each  $d$  and monic linear pencil  $\tilde{L}$  of size  $d$  satisfying  $\mathcal{S}_L \subseteq \mathcal{S}_{\tilde{L}}$ .*

- (2) *If  $\mathcal{D}_L$  is balanced (i.e., for each  $X \in \mathcal{D}_L$  we have  $-X \in \mathcal{D}_L$ ), then for each  $n \in \mathbb{N}$ ,*

$$\tau(L)(n) \geq \frac{1}{n}.$$

*Proof.* The proof appears in Section 8. ■

**1.4. Interpretation in terms of completely positive maps.** The intimate connection between dilations and completely positive maps was first exposted by Stinespring [Sti55] to give abstract necessary and sufficient conditions for the existence of dilations. The theme was explored further by Arveson, see e.g. [Arv69, Arv72]; we refer the reader to [Pau02] for a beautiful exposition. We next explain how our dilation theoretic results pertain to (completely) positive maps.

The equality between the commutability index and the inclusion scale (Theorem 1.4) can be interpreted via positive and completely positive maps. Loosely speaking, each unital positive map can be scaled to a unital completely positive map in a uniform way.

Suppose  $A \in \mathbb{S}_V^g$  is such that the associated spectrahedron  $\mathcal{S}_{L_A}$  is bounded. If  $\tilde{A} \in \mathbb{S}_\eta^g$  is another  $g$ -tuple, consider the unital linear map

$$(1.9) \quad \begin{aligned} \Phi : \text{span}\{I, A_1, \dots, A_g\} &\rightarrow \text{span}\{I, \tilde{A}_1, \dots, \tilde{A}_g\} \\ A_j &\mapsto \tilde{A}_j. \end{aligned}$$

(It is easy to see that  $\Phi$  is well-defined by the boundedness of  $\mathcal{S}_{L_A}$ .) For  $c \in \mathbb{R}$  we define the following scaled distortion of  $\Phi$ :

$$\begin{aligned} \Phi_c : \text{span}\{I, A_1, \dots, A_g\} &\rightarrow \text{span}\{I, \tilde{A}_1, \dots, \tilde{A}_g\} \\ I &\mapsto I \\ A_j &\mapsto c\tilde{A}_j. \end{aligned}$$

**Corollary 1.5.** *With the setup as above,  $c := \tau(L_A)(\eta)$  is the largest scaling factor with the following property: if  $\tilde{A} \in \mathbb{S}_\eta^g$  and if  $\Phi$  is positive, then  $\Phi_c$  completely positive.*

*Proof.* A map  $\Phi$  as in (1.9) is  $k$ -positive iff  $\mathcal{D}_{L_A}(k) \subseteq \mathcal{D}_{L_{\tilde{A}}}(k)$  by [HKM13, Theorem 3.5]. Now apply Theorem 1.4. ■



**1.5. Matrix cube problem.** Given  $A \in \mathbb{S}_d^g$ , the matrix cube problem of Ben-Tal and Nemirovski [B-TN02] is to determine, whether  $\mathcal{S}_{L_C} = [-1, 1]^g \subseteq \mathcal{S}_{L_A}$ . While their primary interest in this problem came from robust analysis (semidefinite programming with interval uncertainty and quadratic Lyapunov stability analysis and synthesis), this problem is in fact far-reaching. For instance, [B-TN02] shows that determining the maximum of a positive definite quadratic form over the unit cube is a special case of the matrix cube problem (cf. Nesterov's  $\frac{\pi}{2}$ -Theorem [Nes97], or the Goemans-Williamson [GoW195] SDP relaxation of the Max-Cut problem). Furthermore, it implies the symmetric Grothendieck inequality. A very recent occurrence of the matrix cube problem is described in [BGKP+], see their Equation (1.3). There a problem in statistics is shown equivalent to whether a LMI, whose coefficients are Hadamard products of a given matrix, contains a cube.

Of course, one could test the inclusion  $\mathcal{S}_{L_C} \subseteq \mathcal{S}_{L_A}$  by checking if all vertices of the cube  $\mathcal{S}_{L_C}$  are contained in  $\mathcal{S}_{L_A}$ . However, the number of vertices grows exponentially with the dimension  $g$ . Indeed the matrix cube problem is NP-hard [B-TN02, Nem06]; see also [KTT13]. A principal result in [B-TN02] is the identification of a computable error bound for a natural relaxation of the matrix cube problem. In [HKM13] we observed this relaxation is in fact equivalent to the free relaxation  $\mathfrak{C}^{(g)} \subseteq \mathcal{D}_{L_A}$ .

With the notations introduced above, we can now present the theorem of Ben-Tal and Nemirovski bounding the error of the free relaxation.

**Theorem 1.6** ([B-TN02]). *Given  $g, \nu \in \mathbb{N}$  and  $B \in \mathbb{S}_\nu^g$ , if  $[-1, 1]^g \subseteq \mathcal{S}_{L_B}$ , then*

- (a)  $\mathfrak{C}^{(g)} \subseteq \vartheta(\nu) \mathcal{D}_{L_B}$ ;
- (b) *if  $d \in \mathbb{N}$  is an upper bound for the ranks of the  $B_j$ , then  $\mathfrak{C}^{(g)} \subseteq \vartheta(d) \mathcal{D}_{L_B}$ .*

*Proof.* Part (a) of Theorem 1.6 is shown, in Theorem 5.10, to be a consequence of our Dilation Theorem 1.1. A further argument, carried out in Section 7, establishes part (b). ■

In this article we show that the bound  $\vartheta(d)$  in Theorem 1.6(a) (and hence in (b)) is sharp.

**Theorem 1.7.** *Suppose  $d \in \mathbb{N}$  and  $\vartheta' \in \mathbb{R}$ . If  $1 \leq \vartheta' < \vartheta(d)$ , then there is  $g \in \mathbb{N}$  and  $A \in \mathbb{S}_d^g$  such that  $[-1, 1]^g \subseteq \mathcal{S}_{L_A}$ , but  $\mathfrak{C}^{(g)}(d) \not\subseteq \vartheta' \mathcal{D}_{L_A}(d)$ .*

*Proof.* See Section 5.4. ■

**Remark 1.8.** Theorem 1.4 applied to the free cube(s) implies, for a given  $g$ , that  $\tau(\mathfrak{C}^{(g)})$  equals  $r(\mathfrak{C}^{(g)})$  and, for fixed  $d$ , the sequences  $(\tau(\mathfrak{C}^{(g)})(d))_g$  and  $(r(\mathfrak{C}^{(g)})(d))_g$  termwise decrease with  $g$  to a common limit, which, in view of Theorems 1.6 and 1.7, turns out to be  $\vartheta(d)$ . In particular, for any  $g$  and any  $g$ -tuple  $C$  of symmetric contractive matrices there exists a  $g$ -tuple of commuting self-adjoint contractions  $T$  on Hilbert space such that  $\frac{1}{\vartheta(d)}C$  dilates to  $T$ , a statement considerably weaker than the conclusion of Theorem 1.1. □

**1.6. Matrix balls.** Paulsen [Pau02] studied the family of operator space structures associated to a normed vector space. Among these, he identified canonical maximal and minimal structures and measured the divergence between them with a parameter he denoted by  $\alpha(V)$

[Pau02]. In the case that  $V$  is infinite dimensional,  $\alpha(V) = \infty$  and if  $V$  has dimension  $g$ , then  $\frac{\sqrt{g}}{2} \leq \alpha(V) \leq g$ . Let  $\ell_1^g$  denote the vector space  $\mathbb{C}^g$  with the  $\ell_1$  norm. Its unit ball is the cube  $[-1, 1]^g$  and in this case  $\sqrt{\frac{g}{2}} \leq \alpha(\ell_1^g) \leq \sqrt{g-1}$ .

We pause here to point out differences between his work and the results in this paper, leaving it to the interested reader to consult [Pau02] for precise definitions. Loosely, the maximal and minimal operator space structures involve quantifying over matrices, not necessarily symmetric, of all sizes  $d$ . By contrast, in this article the matrices are symmetric, the coefficient are real, we study operator systems (as opposed to spaces) determined by linear matrix inequalities and, most significantly, to this point the size  $d$  is fixed. In particular, for the matrix cube, the symmetric matrix version of  $\ell_1^g$  with the minimal operator space structure, the parameter  $\vartheta(d)$  obtained by fixing  $d$  and quantifying over  $g$  remains finite.

Let  $\ell_2^g$  denote the  $\mathbb{C}^g$  with the  $\ell_2$  (Euclidean) norm and let  $\mathbb{B}_g$  denote the (Euclidean) unit ball in  $\mathbb{R}^g$ . In Section 14, we consider the free relaxation of the problem of including  $\mathbb{B}_g$  into a spectrahedron with  $g$ , but not  $d$ , fixed. Thus we study the symmetric variable analog of  $\alpha(\ell_2^g)$ . Among our findings is that the worst case inclusion scale is exactly  $g$ . By contrast,  $\alpha(\ell_2^g)$  is only known to be bounded above by  $\frac{g}{2^{\frac{1}{4}}}$  [Pis03] and below by roughly  $\frac{g+1}{2}$  [Pau02].

**1.7. Adapting the Theory to Free Nonsymmetric Variables.** In this brief subsection we explain how our dilation theoretic results extend to nonsymmetric variables. That is, we present the Simultaneous Dilation Theorem (Corollary 1.9) dilating arbitrary contractive complex matrices to commuting normal contractions up to a scaling factor.

**Corollary 1.9** (Simultaneous Dilation). *Let  $d \in \mathbb{N}$ . There is a Hilbert space  $\mathcal{H}$ , a family  $\mathcal{N}_d$  of commuting normal contractions on  $\mathcal{H}$ , and an isometry  $V : \mathbb{C}^d \rightarrow \mathcal{H}$  such that for each complex  $d \times d$  contraction matrix  $X$  there exists a  $T \in \mathcal{N}_d$  such that*

$$\frac{1}{\sqrt{2} \vartheta(2d)} X = V^* T V.$$

*Proof.* Given any  $d \times d$  complex contraction  $X$ , the matrices

$$S = \frac{1}{2}(X + X^*), \quad K = \frac{1}{2i}(X - X^*)$$

are self-adjoint contractions with  $S + iK = X$ . Consider the  $\mathbb{R}$ -algebra  $*$ -homomorphism

$$\iota : M_d(\mathbb{C}) \rightarrow M_{2d}(\mathbb{R})$$

$$(a_{ij})_{i,j=1}^d \mapsto \begin{pmatrix} (\operatorname{Re} a_{ij})_{i,j=1}^d & (\operatorname{Im} a_{ij})_{i,j=1}^d \\ -(\operatorname{Im} a_{ij})_{i,j=1}^d & (\operatorname{Re} a_{ij})_{i,j=1}^d \end{pmatrix}$$

A straightforward calculation shows

$$Y = \frac{1}{2} \begin{pmatrix} I_d \\ iI_d \end{pmatrix}^* \iota(Y) \begin{pmatrix} I_d \\ iI_d \end{pmatrix}$$

for any  $Y \in M_d(\mathbb{C})$ .

Let  $\mathcal{C}_{2d}$ ,  $\mathcal{H}$  and  $V$  be as in Theorem 1.1. Then  $\iota(S), \iota(K)$  dilate up to a factor  $\frac{1}{\vartheta(2d)}$  to elements  $\widehat{S}, \widehat{K}$  of  $\mathcal{C}_{2d}$ . Thus

$$\frac{1}{\vartheta(2d)}(\iota(S) + i\iota(K)) = V^*(\widehat{S} + i\widehat{K})V,$$

whence

$$\frac{1}{\vartheta(2d)}X = W^*(\widehat{S} + i\widehat{K})W,$$

where  $W$  is the isometry

$$W = \frac{1}{\sqrt{2}}V \begin{pmatrix} I_d \\ iI_d \end{pmatrix}.$$

Hence letting  $\mathcal{N}_{2d} = \frac{1}{\sqrt{2}}(\mathcal{C}_{2d} + i\mathcal{C}_{2d})$  we have

$$\frac{1}{\sqrt{2}\vartheta(2d)}X = V^*NV$$

for

$$N = \frac{1}{\sqrt{2}}(\widehat{S} + i\widehat{K}) \in \mathcal{C}_d.$$

It is clear that elements of  $\mathcal{N}_{2d}$  are pairwise commuting normal contractions. ■

**1.8. Probabilistic theorems and interpretations.** This section assumes only basic knowledge about the Binomial and Beta distributions and does not depend upon the rest of this introduction. The proof of Theorem 1.2 produced, as byproducts, several theorems on the Binomial and Beta distribution which are discussed here and in more detail in Section 15.

We thank Ian Abramson for describing a Bayesian perspective.

**1.8.1. Binomial distributions.** With  $\mathfrak{d}$  fixed, perform  $\mathfrak{d}$  independent flips of a biased coin whose probability of coming up heads is  $p$ . Let  $\mathfrak{S}$  denote the random variable representing the number of heads which occur, and let  $P_p(\mathfrak{S} = \mathfrak{s})$  denote the probability of getting exactly  $\mathfrak{s}$  heads. On the same probability space is the random variable  $\mathfrak{T}$  which represents the number of tails which occur. Of course  $\mathfrak{T} = \mathfrak{d} - \mathfrak{S}$  and the probability of getting exactly  $\mathfrak{t}$  tails is denoted  $P_p(\mathfrak{T} = \mathfrak{t})$ . The distribution of  $\mathfrak{S}$ ,

$$\text{Bin}(\mathfrak{d}, p; \mathfrak{s}) := \binom{\mathfrak{d}}{\mathfrak{s}} p^{\mathfrak{s}} (1-p)^{\mathfrak{t}},$$

at  $\mathfrak{s}$  is **Binomial** with parameters  $p$  and  $\mathfrak{d}$ . Our main interest will be behavior of functions of the form  $P_{p(\mathfrak{s})}(\mathfrak{S} \geq \mathfrak{s})$  for a function  $p(\mathfrak{s})$  close to  $\frac{\mathfrak{s}}{\mathfrak{d}}$ .

The Cumulative Distribution Function (CDF) of a **Beta Distributed** random variable  $\mathfrak{B}$  with shape parameters  $\mathfrak{s}, \mathfrak{t}$  is the function of  $x$  denoted  $P^{b(\mathfrak{s}, \mathfrak{t})}(\mathfrak{B} \leq x) = I_x(\mathfrak{s}, \mathfrak{t})$ . Its mean is  $\frac{\mathfrak{s}}{\mathfrak{s} + \mathfrak{t}}$  and its probability density function (PDF) is

$$\varrho_{\mathfrak{s}, \mathfrak{t}}(x) = \frac{1}{B(\mathfrak{s}, \mathfrak{t})} x^{\mathfrak{s}-1} (1-x)^{\mathfrak{t}-1}.$$

When  $\mathfrak{s}, \mathfrak{t}$  are integers, The CDF for the Binomial Distribution with  $\mathfrak{s} + \mathfrak{t} = \mathfrak{d}$  can be recovered via

$$(1.10) \quad P_p(\mathfrak{S} \geq \mathfrak{s}) = I_p(\mathfrak{s}, \mathfrak{t} + 1) = P^{b(\mathfrak{s}, \mathfrak{t}+1)}(\mathfrak{B} \leq p).$$

For the complementary CDF using  $1 - P_p(\mathfrak{S} \geq \mathfrak{s}) = 1 - P^{b(\mathfrak{s}, \mathfrak{t}+1)}(\mathfrak{B} \leq p)$  gives

$$(1.11) \quad P_p(\mathfrak{S} \leq \mathfrak{s} - 1) = P_p(\mathfrak{S} < \mathfrak{s}) = P^{b(\mathfrak{s}, \mathfrak{t}+1)}(\mathfrak{B} \geq p).$$

**1.8.2. Equipoints and medians.** Our results depend on the nature and estimates of medians, means or **equipoints** (defined below) all being measures of central tendency. Recall for any random variable a **median** is defined to be an  $\mathfrak{s}$  in the sample space satisfying  $P(\mathfrak{S} \leq \mathfrak{s}) \geq \frac{1}{2}$  and  $P(\mathfrak{S} \geq \mathfrak{s}) \geq \frac{1}{2}$ . For a Binomial distributed random variable  $P_{\frac{\mathfrak{s}}{\mathfrak{d}}}$  the median is the mean is the mode is  $\mathfrak{s}$  when  $\mathfrak{s}, \mathfrak{d}$  are integers.

Given a binomially distributed random variable  $\mathfrak{S}$ , we call  $e_{\mathfrak{s}, \mathfrak{t}} \in [0, 1]$  an **equipoint of  $\mathfrak{s}$** , provided

$$(1.12) \quad P_{e_{\mathfrak{s}, \mathfrak{t}}}(\mathfrak{S} \geq \mathfrak{s}) = P_{e_{\mathfrak{s}, \mathfrak{t}}}(\mathfrak{S} \leq \mathfrak{s}).$$

Here  $\mathfrak{s}, \mathfrak{t} \in \mathbb{N}$ , and  $\mathfrak{d} = \mathfrak{s} + \mathfrak{t}$ . Since  $P_{e_{\mathfrak{s}, \mathfrak{t}}}(\mathfrak{S} \geq \mathfrak{s}) + P_{e_{\mathfrak{s}, \mathfrak{t}}}(\mathfrak{S} \leq \mathfrak{s}) \geq 1$ , Equation (1.12) implies  $\mathfrak{s}$  is a median for  $\mathfrak{S}$ . A median is in  $\mathbb{N}$  for Binomial and in  $\mathbb{R}$  for Beta distributed random variables. In practice equipoints and means are close. For example, when the PDF is  $\text{Bin}(10, \frac{5}{10})$  the mean is  $\frac{5}{10}$  one can compute  $e_{\mathfrak{s}, \mathfrak{t}}$  for  $\mathfrak{s} = 1, \dots, 10$ :

$\mathfrak{s}$	1	2	3	4	5	6	7	8	9	10
$e_{\mathfrak{s}, 10-\mathfrak{s}}$	0.111223	0.208955	0.306089	0.403069	0.5	0.596931	0.693911	0.791045	0.888777	1

In contrast, the Beta Distribution is continuous, so for  $\mathfrak{s}, \mathfrak{t} \in \mathbb{R}_{\geq 0}$  with  $\mathfrak{d} = \mathfrak{s} + \mathfrak{t} > 0$  we define  $e_{\mathfrak{s}, \mathfrak{t}}$  by

$$(1.13) \quad P^{b(\mathfrak{s}+1, \mathfrak{t})}(\mathfrak{B} \leq e_{\mathfrak{s}, \mathfrak{t}}) = P^{b(\mathfrak{s}, \mathfrak{t}+1)}(\mathfrak{B} \geq e_{\mathfrak{s}, \mathfrak{t}}),$$

and we call  $e_{\mathfrak{s}, \mathfrak{t}}$  the **equipoint** of the  $\text{Beta}(\mathfrak{s}, \mathfrak{t})$  distribution. Equivalently,

$$P^{b(\mathfrak{s}, \mathfrak{t}+1)}(\mathfrak{B} \leq e_{\mathfrak{s}, \mathfrak{t}}) + P^{b(\mathfrak{s}+1, \mathfrak{t})}(\mathfrak{B} \leq e_{\mathfrak{s}, \mathfrak{t}}) = 1.$$

In terms of the regularized beta function,  $e_{\mathfrak{s}, \mathfrak{t}}$  is determined by

$$(1.14) \quad I_{e_{\mathfrak{s}, \mathfrak{t}}}(s, t + 1) + I_{e_{\mathfrak{s}, \mathfrak{t}}}(s + 1, t) = 1.$$

When  $\mathfrak{s}, \mathfrak{t}$  are integers, the probabilities in (1.12) and (1.13) coincide, so the two definitions give the same  $e_{\mathfrak{s}, \mathfrak{t}}$ . Verifying this statement is an exercise in the notations. The connection between equipoints and the theory of the matrix cube emerges in Section 10.

**Example 1.10.** Here is a concrete probabilistic interpretation of the equipoint  $e_{\mathfrak{s}, \mathfrak{t}}$ . Joe flips a biased coin with probability  $p$  of coming up heads, but does not know  $p$ . After  $\mathfrak{s} - 1$  heads and  $\mathfrak{t} - 1$  tails, the probability that  $p$  is less than  $r$  is  $I_r(\mathfrak{s}, \mathfrak{t})$  by Bayes' Theorem<sup>1</sup>.

The equipoint  $e_{\mathfrak{s}, \mathfrak{t}}$  pertains to the next toss of the coin. If it is a head (resp. tail), then  $b(\mathfrak{s} + 1, \mathfrak{t})$  (resp.  $b(\mathfrak{s}, \mathfrak{t} + 1)$ ) becomes the new distribution for estimating  $p$ . From (1.13), the

<sup>1</sup>[https://en.wikipedia.org/wiki/Checking\\_whether\\_a\\_coin\\_is\\_fair](https://en.wikipedia.org/wiki/Checking_whether_a_coin_is_fair)

equipoint is defined so that with the next toss *the probability after a head that  $p$  is at most  $e_{\mathfrak{s},\mathfrak{t}}$  equals the probability after a tail that  $p$  is at least  $e_{\mathfrak{s},\mathfrak{t}}$* .  $\square$

The next two subsections contain more information on equipoints and how they compare to means and medians.

**1.8.3. Equipoints compared to medians.** Here is a basic property of equipoints versus medians and means. Let  $\frac{1}{2}\mathbb{N}$  denote the set of all positive half-integers, i.e., all  $\mathfrak{s} = \frac{s}{2}$  with  $s \in \mathbb{N}$ .

**Theorem 1.11.** *For  $\mathfrak{d} \in \frac{1}{2}\mathbb{N}$  and  $\mathfrak{s}, \mathfrak{t} \in \frac{1}{2}\mathbb{N}$  with  $\mathfrak{s} + \mathfrak{t} = \mathfrak{d}$ , if  $\frac{\mathfrak{d}}{2} \leq s < \mathfrak{d}$ , then*

$$(1.15) \quad P_{\frac{\mathfrak{s}}{\mathfrak{d}}}(\mathfrak{S} < \mathfrak{s}) \leq P_{\frac{\mathfrak{s}}{\mathfrak{d}}}(\mathfrak{S} > \mathfrak{s}) \quad \text{provided } \mathfrak{s}, \mathfrak{t}, \mathfrak{d} \in \mathbb{N};$$

$$(1.16) \quad P^{b(\mathfrak{s}, \mathfrak{t}+1)}\left(\mathfrak{B} \geq \frac{\mathfrak{s}}{\mathfrak{d}}\right) \leq P^{b(\mathfrak{s}+1, \mathfrak{t})}\left(\mathfrak{B} \leq \frac{\mathfrak{s}}{\mathfrak{d}}\right).$$

Both (1.15) and (1.16) are equivalent to

$$(1.17) \quad e_{\mathfrak{s}, \mathfrak{t}} \leq \frac{\mathfrak{s}}{\mathfrak{d}}.$$

We also have the lower bound

$$(1.18) \quad \frac{\mathfrak{s} + 1}{\mathfrak{s} + \mathfrak{t} + 2} \leq e_{\mathfrak{s}, \mathfrak{t}},$$

for real numbers  $\mathfrak{s} \geq \mathfrak{t} \geq 1$ .

**Remark 1.12.** The inequality (1.16) for integer  $\mathfrak{s}, \mathfrak{t}$  is Simmons' Theorem, cf. [PR07]. The lower bound is new. For half-integer  $\mathfrak{s}, \mathfrak{t}$  both our upper and lower bounds are new.  $\square$

*Proof.* The inequality (1.16) implies (1.15) by (1.10) and (1.11). However, (1.16) and  $e_{\mathfrak{s}, \mathfrak{t}} \leq \frac{\mathfrak{s}}{\mathfrak{d}}$  is the content of Theorem 10.1. The lower bound (1.18) is Proposition 11.2.  $\blacksquare$

Computer experiments lead us to believe (1.17) is true for real numbers:

**Conjecture 1.13.** For  $\mathfrak{s}, \mathfrak{t} \in \mathbb{R}_{>0}$  with  $\mathfrak{s} \geq \mathfrak{t}$ , inequality (1.16) holds. Equivalently,  $e_{\mathfrak{s}, \mathfrak{t}} \leq \frac{\mathfrak{s}}{\mathfrak{s} + \mathfrak{t}}$ .

As a side-product of our quest for bounds on the equipoint we obtain new upper bounds on the median  $m_{\alpha, \beta}$  of the Beta Distribution  $\text{Beta}(\alpha, \beta)$ .

**Corollary 1.14.** *Suppose  $\mathfrak{s}, \mathfrak{t} \in \mathbb{R}$ . If  $1 \leq \mathfrak{s} \leq \mathfrak{t}$  and  $\mathfrak{s} + \mathfrak{t} \geq 3$ , then*

$$\mu_{\mathfrak{s}, \mathfrak{t}} := \frac{\mathfrak{s}}{\mathfrak{s} + \mathfrak{t}} \leq m_{\mathfrak{s}, \mathfrak{t}} \leq \mu_{\mathfrak{s}, \mathfrak{t}} + \frac{\mathfrak{s} - \mathfrak{t}}{(\mathfrak{s} + \mathfrak{t})^2}.$$

*Proof.* The lower bound is known, see [GM77, PYY89]. The upper bound is proved in Corollary 11.7.  $\blacksquare$

1.8.4. *Monotonicity of the CDF.* A property of the functions

$$(1.19) \quad \Phi(\mathfrak{s}) := P^{b(\mathfrak{s}, \mathfrak{d}-\mathfrak{s}+1)}(\mathfrak{B} \leq e_{\mathfrak{s}, \mathfrak{d}-\mathfrak{s}}) \quad \text{and} \quad \hat{\Phi}(\mathfrak{s}) := P^{b(\mathfrak{s}, \mathfrak{d}-\mathfrak{s}+1)}\left(\mathfrak{B} \leq \frac{\mathfrak{s}}{\mathfrak{d}}\right),$$

where  $\mathfrak{B}$  is a Beta distributed random variable, is *one step monotonicity*.

**Theorem 1.15.** *Fix  $0 < \mathfrak{d} \in \mathbb{R}$ .*

- (1)  $\hat{\Phi}(\mathfrak{s}) \leq \hat{\Phi}(\mathfrak{s} + 1)$  for  $\mathfrak{s} \in \mathbb{R}$  with  $\frac{\mathfrak{d}}{2} \leq \mathfrak{s} < \mathfrak{d} - 1$ ;
- (2)  $\Phi(\mathfrak{s}) \leq \Phi(\mathfrak{s} + 1)$  for  $\mathfrak{s}, \mathfrak{d} \in \frac{1}{2}\mathbb{N}$  with  $\frac{\mathfrak{d}}{2} \leq \mathfrak{s} < \mathfrak{d} - 1$ .

*Proof.* See Section 15.2. ■

Computer experiments lead us to believe monotonicity of  $\Phi$  holds for real numbers  $\mathfrak{s}, \mathfrak{t}$ .

**Conjecture 1.16.**  $\Phi(\mathfrak{s}) < \Phi(\tilde{\mathfrak{s}})$  for  $\mathfrak{s}, \tilde{\mathfrak{s}}, \mathfrak{d} \in \mathbb{R}$  with  $0 < \frac{\mathfrak{d}}{2} \leq \mathfrak{s} < \tilde{\mathfrak{s}} < \mathfrak{d}$ .

The functions  $\Phi$  and  $\hat{\Phi}$  are based on the CDF. Analogous results hold for the PDF and these appear in Section 15.2.1.

The monotonicity result of Theorem 1.15 allows us to identify the minimizers of  $\Phi$ . Indeed the following theorem restates Theorem 1.2 in probabilistic terms.

**Theorem 1.17.** *For  $\mathfrak{d} \in \frac{1}{2}\mathbb{N}$  and  $\frac{\mathfrak{d}}{2} \leq \mathfrak{s} < \mathfrak{d} - 1$ , the function  $\Phi$  of  $\mathfrak{s} \in \frac{1}{2}\mathbb{N}$  takes its minimum at*

- (1)  $\mathfrak{s} = \mathfrak{t} = \frac{\mathfrak{d}}{2}$  if  $\mathfrak{d} \in \mathbb{N}$ ;
- (2)  $\mathfrak{s} = \mathfrak{d} + \frac{1}{2}$  and  $\mathfrak{t} = \mathfrak{d} - \frac{1}{2}$  if  $\mathfrak{d} \in \frac{1}{2}\mathbb{N} \setminus \mathbb{N}$ .

**1.9. Reader's guide.** The rest of this article is organized as follows. Results relating dilations to free spectrahedral inclusions needed for the proofs of the results for the matrix cube are collected in Section 2. Further general results on free spectrahedral inclusions and dilations appear in Section 8. The results of Section 4 simplify the identification of the optimum  $\vartheta(d)$  as defined by Equation (1.1). They also identify, implicitly, constrained versions of this optimum. The results of the previous sections are combined in Section 5 to prove Theorem 1.1 and Theorem 1.7, as well as the weaker version of Theorem 1.6 which asks that the matrices  $B$  have size  $d$ , and not just rank at most  $d$ . In Section 6, the constrained optima from Section 4 are identified, still implicitly, in terms of the regularized incomplete beta function, a result needed to complete the proof of Theorem 1.6 in Section 7 as well as in the remaining sections of the paper. Theorem 1.2 is reformulated in terms of the beta function in advance of the following three sections which together establish Theorem 1.2. A half-integer generalization of Simmons' Theorem, inspired by the strategy in [PR07] using two step monotonicity, is the topic of Section 10. A new lower bound for the median of the Beta distribution and bounds for the equipoint appear in Section 11 and the bounds for the equipoint are used in Section 12 to complete the proof of Theorem 1.2. Estimates for  $\vartheta(d)$  in the case that  $d$  is odd appear in Section 13. Section 14 considers the problem of including the unit ball in  $\mathbb{R}^g$  into a spectrahedron, and

uses dilation theory to prove the worst case error inherent in its free relaxation, namely  $g$ . Finally, further probabilistic results and their proofs are expositied in Section 15.

The reader interested only in probabilistic results can proceed to Sections 10, 11 and 15. The reader interested only in the matrix cube problem can skip Section 8; whereas the reader interested only in dilation results (absent formulas for  $\vartheta(d)$ ) can focus on the sections up through and including Section 8.

The original version of this manuscript<sup>2</sup> treated the case where the rank (size)  $d$  of the LMI defining pencil of the containing spectrahedron is fixed, but the number of variables  $g$  is not. Subsequently, we and Davidson, Dor-On, Shalit and Solel independently consider inclusion problems for balls (where  $g$  is fixed, but  $d$  is not). Our results in this direction appear in Section 14. We highly recommend the posting [DDSS+] for its many interesting results, including a far reaching generalization of the symmetry based inclusion result of Proposition 14.1 and a fascinating connection with the theory of frames. In addition [DDSS+] extends many spectrahedral inclusion results to the setting of (infinite dimensional) operators. (See also [Zal+] which considers the operator setting with an emphasis on the unbounded domains as well as various polynomial Positivstellensätze.) They also show if a tuple  $X$  of matrices dilates to a commuting tuple  $N$  of normal operators, then  $X$  dilates to a commuting tuple of normal matrices  $T$  satisfying the inclusion  $\sigma(T) \subseteq \sigma(N)$ . In particular, using the (infinite dimensional) optimum bound  $\vartheta(d)$  identified in this article, it follows that given a tuple  $X$  (finite set) of  $d \times d$  symmetric contractions, there exists a tuple  $T$  of commuting symmetric matrices and an isometry  $V$  such that  $X = \vartheta(d)V^*TV$ .

## 2. DILATIONS AND FREE SPECTRAHEDRAL INCLUSIONS

This section presents preliminaries on free spectrahedral inclusions and dilations, tying the existence of dilations to appropriate commuting tuples (the commutability index) to free spectrahedral inclusion.

The following proposition gives a sufficient condition for the inclusion of one spectrahedron in another. It will later be applied to  $\mathfrak{C}$ , the free cube. Recall the definitions of  $L_A(x)$ ,  $L_A(X)$ ,  $\mathcal{D}_{L_A}$  and  $\mathcal{S}_{L_A}$  from Section 1.3. Any  $r > 0$  (and necessarily  $1 \geq r$ ) with the property that the inclusion  $\mathcal{S}_{L_A} \subseteq \mathcal{S}_{L_B}$  implies the inclusion  $r\mathcal{D}_{L_A} \subseteq \mathcal{D}_{L_B}$  provides an estimate for the error in testing spectrahedral inclusion using the free spectrahedral inclusion as a relaxation. Indeed, suppose that  $\mathcal{S}_{L_A} \subseteq \mathcal{S}_{L_B}$ , but, for  $t > 1$ , that  $t\mathcal{S}_{L_A} \not\subseteq \mathcal{S}_{L_B}$ . The free relaxation amounts to finding the largest  $\rho$  such that  $\rho\mathcal{D}_{L_A} \subseteq \mathcal{D}_{L_B}$  and concluding that necessarily  $\rho\mathcal{S}_{L_A} \subseteq \mathcal{S}_{L_B}$ . Since  $\rho \geq r$ , it follows that  $r$  then provides a lower bound for the error.

**Proposition 2.1.** *Suppose  $A$  is a  $g$ -tuple of symmetric  $m \times m$  matrices,  $d$  is a positive integer,  $r > 0$  and for each  $X \in \mathcal{D}_{L_A}(d)$  there is a Hilbert space  $\mathcal{H}$ , an isometry  $W: \mathbb{R}^d \rightarrow \mathcal{H}$  and a tuple  $T = (T_1, \dots, T_g)$  of commuting bounded self-adjoint operators  $T_i$  on  $\mathcal{H}$  with joint spectrum contained in  $\mathcal{S}_{L_A}$  such that  $rX_i = W^*T_iW$  for all  $i \in \{1, \dots, g\}$ . If  $B$  is a tuple of  $d \times d$  symmetric matrices and  $\mathcal{S}_{L_A} \subseteq \mathcal{S}_{L_B}$ , then  $r\mathcal{D}_{L_A} \subseteq \mathcal{D}_{L_B}$ .*

---

<sup>2</sup><https://arxiv.org/abs/1412.1481>

**Remark 2.2.** It turns out that in the case of  $\mathcal{D}_{L_A} = \mathfrak{C}^{(g)}$ , one only needs to assume that  $B$  is a tuple of  $m \times m$  symmetric matrices each of rank at most  $d$ . See [B-TN02] or Section 7 of this paper, where an elaboration on the argument below plus special properties of the  $\mathfrak{C}^{(g)}$  are used to establish this result.  $\square$

The proof of the proposition employs the following lemma which will also be used in Section 7.

**Lemma 2.3.** *If  $A$  is a  $g$ -tuple of symmetric  $m \times m$  matrices,  $B$  is a  $g$ -tuple of symmetric  $d \times d$  matrices and if  $\varrho\mathcal{D}_{L_A}(d) \subseteq \mathcal{D}_{L_B}(d)$ , then  $\varrho\mathcal{D}_{L_A}(n) \subseteq \mathcal{D}_{L_B}(n)$  for every  $n$ .*

*Proof.* Suppose  $\varrho\mathcal{D}_{L_A}(d) \subseteq \mathcal{D}_{L_B}(d)$ ,  $n \in \mathbb{N}$ ,  $n \geq d$  and  $(X_1, \dots, X_g) \in \mathcal{D}_{L_A}(n)$ . We have to show that  $\varrho(X_1, \dots, X_g) \in \mathcal{D}_{L_B}(n)$ . Given  $x \in \mathbb{R}^d \otimes \mathbb{R}^n$ , write

$$x = \sum_{s=1}^d e_s \otimes x_s,$$

where  $e_s$  are the standard basis vectors of  $\mathbb{R}^d$ . Let  $M$  denote the span of  $\{x_s : 1 \leq s \leq d\}$  and let  $P$  denote the projection onto  $M$ . Then

$$\begin{aligned} \langle (\sum B_j \otimes X_j)x, x \rangle &= \sum_{j,s,t} \langle B_j e_s, e_t \rangle \langle X_j x_s, x_t \rangle \\ (2.1) \quad &= \sum_{j,s,t} \langle B_j e_s, e_t \rangle \langle P X_j P x_s, x_t \rangle \\ &= \langle (\sum B_j \otimes P X_j P)x, x \rangle. \end{aligned}$$

Now  $\varrho P X P = \varrho(P X_1 P, \dots, P X_g P) \in \varrho\mathcal{D}_{L_A}(r) \subseteq \mathcal{D}_{L_B}(r)$  where  $r \leq d$  is the dimension of  $M$ . Hence, by Equation (2.1),

$$\varrho \langle (\sum B_j \otimes X_j)x, x \rangle \leq \|x\|^2.$$

For  $n < d$ , simply taking a direct sum with 0 (of size  $n - d$ ) produces the tuple  $X \oplus 0 \in \mathcal{D}_{L_A}(d)$  and hence  $\varrho X \oplus 0 \in \mathcal{D}_{L_B}(d)$  by hypothesis. Compressing to the first summand gives  $\varrho X \in \mathcal{D}_{L_B}(n)$  and the proof is complete.  $\blacksquare$

*Proof of Proposition 2.1.* Suppose  $\mathcal{S}_{L_A} \subseteq \mathcal{S}_{L_B}$  and let  $X \in \mathcal{D}_{L_A}(d)$ . Choose a Hilbert space  $\mathcal{H}$ , an isometry  $W: \mathbb{R}^d \rightarrow \mathcal{H}$  and a tuple  $T = (T_1, \dots, T_g)$  of commuting bounded self-adjoint operators  $T_i$  on  $\mathcal{H}$  with joint spectrum contained in  $\mathcal{S}_{L_A}$  such that  $r X_i = W^* T_i W$  for all  $i \in \{1, \dots, g\}$ . Then the joint spectrum of  $T$  is contained in  $\mathcal{S}_{L_B}$ . Writing  $B = \sum_{i=1}^g B_i x_i$  with symmetric  $B_i \in \mathbb{S}_n$ , we have to show that  $r \sum_{i=1}^g B_i \otimes X_i \preceq I_{dn}$ . Let  $E$  denote the joint



spectral measure of  $T$  whose support is contained in  $\mathcal{S}_{L_B} \subseteq \mathbb{R}^g$ . Then

$$\begin{aligned}
r \sum_{i=1}^g B_i \otimes X_i &= \sum_{i=1}^g B_i \otimes W^* T_i W \\
&= \sum_{i=1}^g B_i \otimes W^* \left( \int_{\mathcal{S}_{L_B}} y_i dE(y) \right) W \\
&= \int_{\mathcal{S}_{L_B}} \underbrace{\left( \sum_{i=1}^g B_i y_i \right)}_{\preceq I_d} \otimes \underbrace{W^* dE(y) W}_{\succeq 0} \\
&\preceq \int_{\mathcal{S}_{L_B}} I_d \otimes W^* dE(y) W \\
&= I_d \otimes W^* \left( \int_{\mathcal{S}_{L_B}} dE(y) \right) W \\
&= I_d \otimes W^* \text{id}_{\mathcal{H}} W = I_d \otimes W^* W = I_d \otimes I_n = I_{dn}.
\end{aligned}$$

Hence  $X \in \mathcal{D}_{L_B}(d)$ . An application of Lemma 2.3 completes the proof.  $\blacksquare$

### 3. LIFTING AND AVERAGING

This subsection details the connection between averages of matrices over the orthogonal group and the dilations of tuples of symmetric matrices to commuting tuples of contractive self-adjoint operators, a foundation for the proof of Theorem 1.1.

Let  $M_d$  denote the collection of  $d \times d$  matrices. Let  $O(d) \subseteq M_d$  denote the orthogonal group and let  $dU$  denote the Haar measure on  $O(d)$ . Let  $\mathfrak{D}(d)$  denote the collection of measurable functions  $D: O(d) \rightarrow M_d$  which take diagonal contractive values. Thus, letting  $\mathfrak{M}(O(d), M_d)$  denote the measurable functions from  $O(d)$  to  $M_d$ ,

$$(3.1) \quad \mathfrak{D}(d) = \{D \in \mathfrak{M}(O(d), M_d) : D(U) \text{ is diagonal and } \|D(U)\| \leq 1 \text{ for every } U \in O(d)\}.$$

Let  $\mathcal{H}$  denote the Hilbert space

$$(3.2) \quad \mathcal{H} = \mathbb{R}^d \otimes L^2(O(d)) = L^2(O(d))^d = L^2(O(d), \mathbb{R}^d).$$

Let  $V: \mathbb{R}^d \rightarrow \mathcal{H}$  denote the mapping

$$(3.3) \quad Vx(U) = x.$$

Thus,  $V$  embeds  $\mathbb{R}^d$  into  $\mathcal{H}$  as the constant functions. For  $D \in \mathfrak{D}(d)$ , define  $M_D: \mathcal{H} \rightarrow \mathcal{H}$  by

$$(M_D f)(U) = U D(U) U^* f(U)$$

for all  $U \in O(d)$ . Because  $D(U)$  is pointwise a symmetric contraction for all  $U \in O(d)$ ,  $M_D$  is a self-adjoint contraction on  $\mathcal{H}$ . Let

$$(3.4) \quad \mathcal{C}_d = \{M_D : D \in \mathfrak{D}(d)\}.$$

**Remark 3.1.** Alternately one could define  $V$  by  $Vx(U) = U^*x$  instead of conjugating  $D(U)$  by  $U$  and  $U^*$ .  $\square$

**Lemma 3.2.** *Each  $M_D$  is a self-adjoint contraction.*

**Lemma 3.3.** *If  $D, E \in \mathfrak{D}(d)$ , then  $M_{DE} = M_D \circ M_E = M_E \circ M_D$ . Thus,  $M_D$  and  $M_E$  commute.*

*Proof.* The result follows from the fact that  $D$  and  $E$  pointwise commute and hence the functions  $U \mapsto UD(U)U^*$  and  $U \mapsto UE(U)U^*$  pointwise commute.  $\blacksquare$

**Lemma 3.4.** *The mapping  $V$  is an isometry and its adjoint*

$$V^*: L^2(O(d), \mathbb{R}^d) \rightarrow \mathbb{R}^d$$

*is given by*

$$V^*(f) = \int_{O(d)} f(U) dU$$

*for all  $f \in L^2(O(d), \mathbb{R}^d)$ .*

*Proof.* For all  $x \in \mathbb{R}^n$  and  $f \in L^2(O(d), \mathbb{R}^d)$ , we have

$$\langle Vx, f \rangle = \int_{O(d)} \langle x, f(U) \rangle dU = \left\langle x, \int_{O(d)} f(U) dU \right\rangle,$$

thus computing  $V^*$ . Moreover,  $V$  is an isometry as

$$\langle Vx, Vx \rangle = \left\langle x, \int_{O(d)} x dU \right\rangle = \langle x, x \rangle. \quad \blacksquare$$

**Lemma 3.5.** *For  $D \in \mathfrak{D}(d)$ ,*

$$V^* M_D V = \int_{O(d)} UD(U)U^* dU.$$

For notation purposes, let

$$(3.5) \quad C_D = \int_{O(d)} UD(U)U^* dU.$$

*Proof.* For  $x \in \mathbb{R}^d$ , we have,

$$\begin{aligned}
C_D x &= \left( \int_{O(d)} U D(U) U^* dU \right) x \\
&= \int_{O(d)} U D(U) U^* x dU \\
&= \int_{O(d)} U D(U) U^* ((Vx)(U)) dU \\
&= \int_{O(d)} (M_D(Vx))(U) dU \\
&= V^*(M_D(Vx)). \quad \blacksquare
\end{aligned}$$

**Remark 3.6.** Suppose  $\mathcal{S}$  is a subset of  $\mathfrak{D}(d)$  and consider the family of symmetric matrices  $(C_D)_{D \in \mathcal{S}}$ . The lemmas in this subsection imply that this family dilates to the commuting family of self-adjoint contractions  $(M_D)_{D \in \mathcal{S}}$ . Let  $\mu(D) := \frac{1}{\|C_D\|}$  for  $D \in \mathfrak{D}(d)$  and suppose

$$\mu := \sup\{\mu(D) : D \in \mathcal{S}\}$$

is finite. It follows that the collection of symmetric contractions  $(\mu(D)C_D)_{D \in \mathcal{S}}$  dilates to the commuting family  $(M_{\mu(D)D} = \mu(D)M_D)_{D \in \mathcal{S}}$  of self-adjoint operators of operator norm at most  $\mu$ ,

$$\mu(D)C_D = \mu(D)V^*M_DV$$

for all  $D \in \mathcal{S}$ .

Our aim, in the next few sections, is to turn this construction around. Namely, given a family  $\mathcal{C} \subseteq M_d$  of symmetric contractions (not necessarily commuting), we hope to find a  $t \in [0, 1]$  (as large as possible) and a family  $(F_C)_{C \in \mathcal{C}}$  in  $\mathfrak{D}(d)$  such that

$$tC = V^*M_{F_C}V.$$

for all  $C \in \mathcal{C}$ . Any  $t$  so obtained feeds into the hypotheses of Proposition 2.1.  $\square$

#### 4. A SIMPLIFIED FORM FOR $\vartheta$

Given a symmetric matrix  $B$ , let  $\text{sign}_0(B) = (p, n)$ , where  $p, n \in \mathbb{N}_0$  denote the number of nonnegative and negative eigenvalues of  $B$  respectively. It is valuable to think of the optimization problem (1.1) over symmetric matrices  $B$  in two stages based on the signature. Let

$$(4.1) \quad \kappa_*(s, t) = \min_{\substack{B \in \mathbb{S}_d \\ \text{sign}_0(B) = (s, t) \\ \text{tr } |B| = d}} \int_{S^{d-1}} |\xi^* B \xi| d\xi$$

and note that the minimization (1.1) is

$$(4.2) \quad \frac{1}{\vartheta(d)} = \kappa_*(d) := \min\{\kappa_*(s, t) : s + t = d\}.$$

Given  $s, t \in \mathbb{N}_0$  and  $a, b \geq 0$ , let

$$J(s, t; a, b) := aI_s \oplus -bI_t.$$

Thus  $J(s, t; a, b)$  is the diagonal matrix whose diagonal reads

$$\underbrace{a, \dots, a}_{s \text{ times}}, \underbrace{-b, \dots, -b}_{t \text{ times}}.$$

We simplify the optimization problem (1.1) as follows. The first step consists of showing, for fixed  $s, t \in \mathbb{N}_0$  (with  $s + t = d$ ), that the optimization can be performed over the set  $J(s, t; a, b)$  for  $a, b \geq 0$  such that  $as + bt = d$ . The second step is to establish, again for fixed integers  $s, t$ , an implicit criteria to identify the values of  $a, b$  which optimize (4.1). Toward this end we introduce the following notations. Define

$$(4.3) \quad \kappa(s, t; a, b) := \int_{S^{d-1}} |\xi^* J(s, t; a, b) \xi| d\xi.$$

Finally, let for  $1 \leq j \leq s$ ,

$$(4.4) \quad \alpha(s, t; a, b) = \int_{S^{d-1}} \operatorname{sgn} [\xi^* J(s, t; a, b) \xi] \xi_j^2 d\xi,$$

and, for  $s + 1 \leq j \leq d$ ,

$$(4.5) \quad \beta(s, t; a, b) = - \int_{S^{d-1}} \operatorname{sgn} [\xi^* J(s, t; a, b) \xi] \xi_j^2 d\xi.$$

(It is straightforward to check that  $\alpha$  and  $\beta$  are independent of the choices of  $j$ .)

**Remark 4.1.** The quantities  $\alpha = \alpha(s, t; a, b)$  and  $\beta = \beta(s, t; a, b)$  are interpreted in terms of the regularized beta function in Lemma 6.6.  $\square$

**Proposition 4.2.** For each  $d \in \mathbb{N}$ ,  $s, t \in \mathbb{N}_0$  with  $s + t = d$ , the minimum in Equation (4.1) is achieved at a  $B$  of the form  $J(s, t; a, b)$  where  $a, b > 0$  and  $as + bt = d$ , and

$$\kappa_*(s, t) = \min_{as+bt=d} \int_{S^{d-1}} |\xi^* J(s, t; a, b) \xi| d\xi.$$

Moreover,  $\kappa_*(d, 0) = \kappa_*(0, d) \geq \kappa_*(s, t)$  and for  $s, t \in \mathbb{N}$ , the minimum occurs at a pair  $a(s, t), b(s, t)$  uniquely determined by  $a(s, t), b(s, t) \geq 0$ ,  $sa(s, t) + tb(s, t) = d$  and

$$\alpha(s, t; a(s, t), b(s, t)) = \beta(s, t; a(s, t), b(s, t)),$$

so that  $\kappa_*(s, t) = \kappa(s, t; a(s, t), b(s, t))$ . In particular,

$$\kappa_*(s, t) = d\alpha(s, t; a(s, t), b(s, t)) = d\beta(s, t; a(s, t), b(s, t)).$$

Consequently, the minimum in Equation (1.1) is achieved and is

$$(4.6) \quad \min_{s+t=d} \kappa_*(s, t) = \min_{s+t=d} d \beta(s, t; a(s, t), b(s, t)).$$

It occurs at a  $B$  of the form  $J(s, t; a(s, t), b(s, t))$ .

*Proof.* Let  $\mathcal{T}$  denote the set of symmetric  $d \times d$  matrices  $B$  such that  $\text{tr}(|B|) = d$  and note that  $\mathcal{T}$  is a compact subset of  $\mathbb{S}_d$ . Hence (1.1) is well defined in that the infimum in the definition of  $\vartheta(d)$  is indeed a minimum. Fix a  $B \in \mathcal{T}$  and suppose  $B$  has  $s$  nonnegative eigenvalues and  $t$  negative eigenvalues (hence  $s + t = d$ ). Without loss of generality, assume that  $B$  is diagonal with first  $s$  diagonal entries  $a_1, \dots, a_s$  nonnegative and last  $t$  diagonal entries  $-b_{s+1}, \dots, -b_d$  negative (thus  $b_j > 0$ ). Thus,

$$\int_{S^{d-1}} |\xi^* B \xi| d\xi = \int_{S^{d-1}} \left| \sum_{j=1}^s a_j \xi_j^2 - \sum_{j=s+1}^d b_j \xi_j^2 \right| d\xi.$$

Let  $\Sigma$  denote the subgroup of the group of permutations of size  $n$  which leave invariant the sets  $\{1, \dots, s\}$  and  $\{s+1, \dots, n\}$ . Each  $\sigma \in \Sigma$  gives rise to a permutation matrix  $V_\sigma$ . It is readily checked that

$$\int_{S^{d-1}} |\xi^* V_\sigma^* B V_\sigma \xi| d\xi = \int_{S^{d-1}} \left| \sum_{j=1}^s a_{\sigma(j)} \xi_j^2 - \sum_{j=s+1}^d b_{\sigma(j)} \xi_j^2 \right| d\xi.$$

Let  $N$  denote the cardinality of  $\Sigma$  and note that  $a = \frac{1}{N} \sum_{\sigma \in \Sigma} a_{\sigma(j)}$ ,  $b = \frac{1}{N} \sum_{\sigma \in \Sigma} b_{\sigma(j)}$  are independent of  $j$ . Thus,

$$\frac{1}{N} \sum_{\sigma \in \Sigma} V_\sigma^* B V_\sigma = \sum_{\sigma \in \Sigma} \text{diag} \begin{pmatrix} a_{\sigma(1)} & \dots & a_{\sigma(s)} & -b_{\sigma(s+1)} & \dots & -b_{\sigma(d)} \end{pmatrix} = J(s, t; a, b).$$

Further,  $as + bt = d$  (which depends on averaging over the subgroup  $\Sigma$  rather than the full symmetric group) and hence  $J(s, t; a, b) \in \mathcal{T}$ . Therefore,

$$\begin{aligned} \int_{S^{d-1}} |\xi^* J(s, t; a, b) \xi| d\xi &= \int_{S^{d-1}} \left| \xi^* \left( \frac{1}{N} \sum_{\sigma \in \Sigma} V_\sigma^* B V_\sigma \right) \xi \right| d\xi \\ &\leq \frac{1}{N} \sum_{\sigma \in \Sigma} \int_{S^{d-1}} |\xi^* V_\sigma^* B V_\sigma \xi| d\xi = \frac{1}{N} \sum_{\sigma \in \Sigma} \int_{S^{d-1}} |\xi^* B \xi| d\xi \\ &= \int_{S^{d-1}} |\xi^* B \xi| d\xi. \end{aligned}$$

Thus with  $s, t \geq 0$  and  $a, b \geq 0$

$$\min_{\substack{s+t=d \\ as+bt=d}} \int_{S^{d-1}} |\xi^* J(s, t; a, b) \xi| d\xi \leq \min_{\substack{B \in \mathbb{S}_d \\ \text{tr } |B| = d}} \int_{S^{d-1}} |\xi^* B \xi| d\xi.$$

By compactness of the underlying set for fixed  $d$  the minimum is attained; of course on a diagonal matrix.

Finally to write  $\kappa(s, t; a, b)$  in terms of  $\alpha$  and  $\beta$ , note that

$$\begin{aligned}\kappa(s, t; a, b) &= \int_{S^{d-1}} \operatorname{sgn} [\xi^* J(s, t; a, b) \xi] (\xi^* J(s, t; a, b) \xi) d\xi \\ &= \int_{S^{d-1}} \operatorname{sgn} [\xi^* J(s, t; a, b) \xi] \left( a \sum_{j=1}^s \xi_j^2 - b \sum_{j=s+1}^d \xi_j^2 \right) d\xi \\ &= as\alpha(s, t; a, b) + bt\beta(s, t; a, b).\end{aligned}$$

To this point it has been established that there is a minimizer of the form  $B = J(s, t; a, b)$ . First note that  $\alpha, \beta \leq \frac{1}{d}$ , since, for instance,

$$d\alpha(s, t; a, b) \leq d \int_{S^{d-1}} \xi_j^2 d\xi = \int_{S^{d-1}} \sum_{m=1}^d \xi_m^2 d\xi = \int_{S^{d-1}} d\xi = 1.$$

Hence,

$$\kappa(s, t; a, b) = as\alpha(s, t; a, b) + bt\beta(s, t; a, b) \leq \frac{1}{d}(as + bt) = 1.$$

Moreover, in the case that  $s = d$  and  $t = 0$ , then  $B = I$  and  $\kappa_*(d, 0) = 1$ . Hence,  $\kappa_*(d, 0) \geq \kappa_*(s, t)$ . Turning to the case  $s, t \in \mathbb{N}_0$ , observe if  $b = 0$ , then  $a = \frac{d}{s}$  and  $\kappa(s, t; \frac{d}{s}, 0) = 1$  and similarly,  $\kappa(s, t; 0, \frac{d}{t}) = 1$ . Hence, for such  $s, t$ , the minimum is achieved at a point in the interior of the set  $\{as + bt = d : a, b \geq 0\}$ . Thus, it can be assumed that minimum occurs at  $B = J(s, t; a_*, b_*)$  for some  $a_*, b_* > 0$ .

Any other  $J(s, t, a, b)$ , for  $a, b$  near  $a_*, b_*$ , can be written as

$$J(s, t; a_*, b_*) + \lambda J(s, t; t, -s)$$

(in particular,  $a_* + \lambda t > 0$  as well as  $b_* - \lambda s > 0$ ). By optimality of  $a_*, b_*$ ,

$$\begin{aligned}0 &\leq \int |\xi^* J(s, t; a, b) \xi| d\xi - \int |\xi^* J(s, t; a_*, b_*) \xi| d\xi \\ &= \int (\operatorname{sgn}[\xi^* J(s, t; a, b) \xi] - \operatorname{sgn}[\xi^* J(s, t; a_*, b_*) \xi]) \xi^* J(s, t; a_*, b_*) \xi d\xi \\ &\quad + \lambda \int \operatorname{sgn}[\xi^* J(s, t; a, b) \xi] \xi^* J(s, t; t, -s) \xi d\xi,\end{aligned}$$

where the integrals are over  $S^{d-1}$ . Observe that the integrand of the first integral on the right hand side is always nonpositive and is negative on a set of positive measure. Hence this integral is negative. Hence,

$$(4.7) \quad 0 < \lambda \int \operatorname{sgn}[\xi^* J(s, t; a, b) \xi] \xi^* J(s, t; t, -s) \xi d\xi.$$

Choosing  $\lambda > 0$ , dividing Equation (4.7) by  $\lambda$  and letting  $\lambda$  tend to 0 gives

$$0 \leq \int \operatorname{sgn}[\xi^* J(s, t; a_*, b_*) \xi] \xi^* J(s, t; t, -s) \xi d\xi.$$

On the other hand, choosing  $\lambda < 0$ , dividing by  $\lambda$  and letting  $\lambda$  tend to 0 gives the reverse inequality.

Hence,

$$0 = \int \operatorname{sgn}[\xi^* J(s, t; a_*, b_*) \xi] \xi^* J(s, t; t, -s) \xi d\xi = st (\alpha(s, t; a_*, b_*) - \beta(s, t; a_*, b_*)).$$

Finally, the uniqueness of  $a_*, b_*$  follows from the strict inequality in Equation (4.7), since  $\lambda \neq 0$  corresponds exactly to  $(a, b) \neq (a_*, b_*)$ . We henceforth denote  $(a_*, b_*)$  by  $(a(s, t), b(s, t))$ . In particular,  $(a(s, t), b(s, t))$  is uniquely determined by  $a, b \geq 0$ ,  $as + bt = d$  and  $\alpha(s, t; a, b) = \beta(s, t; a, b)$ .  $\blacksquare$

Note from the proof a limitation of our  $\alpha = \beta$  optimality condition. It was obtained by fixing  $s, t$  and then optimizing over  $a, b \geq 0$ . Thus it sheds no light on subsequent minimization over  $s, t$ . To absorb this information requires many of the subsequent sections of this paper.

## 5. $\vartheta$ IS THE OPTIMAL BOUND

In this section we establish Theorems 1.7 and 1.1. We also state and prove a version of Theorem 1.6 under the assumption that  $B$  is a tuple of  $d \times d$  matrices, rather than the (weaker) assumption that it is a tuple of  $n \times n$  matrices each with rank at most  $d$ . In Section 5.2 we begin to connect, in the spirit of Remark 3.6, the norm of  $C_D$  to that of  $M_D$ . The main results are in Section 5.4.

**5.1. Averages over  $O(d)$  equal averages over  $S^{d-1}$ .** The next trivial lemma allows us to replace certain averages over  $O(d)$  with averages over  $S^{d-1}$ .

**Lemma 5.1.** *Suppose  $\mathcal{B}$  is a Banach space and let  $S^{d-1} \subseteq \mathbb{R}^d$  denote the unit sphere. If*

$$(5.1) \quad f : S^{d-1} \rightarrow \mathcal{B},$$

*is an integrable function and  $\gamma \in S^{d-1}$ , then*

$$(5.2) \quad \int_{S^{d-1}} f(\xi) d\xi = \int_{O(d)} f(U\gamma) dU.$$

*In particular,*

$$\int_{O(d)} f(U\gamma) dU$$

*does not depend on  $\gamma \in S^{d-1}$ .*

We next apply Lemma 5.1 to represent the matrices  $C_D$  defined in Equation (3.5) as integrals over  $S^{d-1}$ . Given  $J \in \mathbb{S}_d$  and choosing an arbitrary unit vector  $\gamma$  in  $\mathbb{R}^d$ , define a matrix  $E_J$  by

$$E_J := \int_{O(d)} \operatorname{sgn}[\gamma^* U^* J U \gamma] U \gamma \gamma^* U^* dU.$$

Note that  $E_J$  depends only on  $J$  but not on  $\gamma$ . In fact,

$$(5.3) \quad E_J = \int_{S^{d-1}} \operatorname{sgn}[\xi^* J \xi] \xi \xi^* d\xi$$

by Lemma 5.1. Given  $J \in \mathbb{S}_d$ , define

$$D_J : O(d) \rightarrow M_d$$

into the diagonal matrices given by

$$(5.4) \quad D_J(U) := \sum_{j=1}^d \operatorname{sgn}[e_j^* U^* J U e_j] e_j e_j^*.$$

In particular  $D_J \in \mathfrak{D}(d)$ , where  $\mathfrak{D}(d)$  is defined in Equation (3.1). Recall, from (3.5), the definition of  $C_D$ . In particular,

$$C_{D_J} = \int_{O(d)} U D_J(U) U^* dU.$$

**Lemma 5.2.** *For  $J \in \mathbb{S}_d$ , independent of  $j$ ,*

$$C_{D_J} = d E_J = \sum_{j=1}^d \int_{O(d)} \operatorname{sgn}[e_j^* U^* J U e_j] e_j e_j^* dU = d \int_{S^{d-1}} \operatorname{sgn}[\xi^* J \xi] \xi \xi^* d\xi.$$

*Further, if  $J \neq 0$ , then  $C_{D_J} \neq 0$ .*

*Proof.* The first statement follows from the definitions and Equation 5.3. For the second statement, observe that

$$\operatorname{tr}(E_J J) = \int_{S^{d-1}} |\xi^* J \xi| d\xi > 0. \quad \blacksquare$$

**Remark 5.3.** The rest of this paper uses averaging over the sphere and not the orthogonal group. But it should be noted that various arguments in Section 3 use integration over the orthogonal group in an essential way.  $\square$

**5.2. Properties of matrices gotten as averages.** This section describes properties of the matrices  $E_J$  defined by Equation (5.3).

**Lemma 5.4.** *If  $U$  is an orthogonal matrix and  $J \in \mathbb{S}_d$ , then*

$$E_{U^* J U} = U^* E_J U.$$

*Proof.* Using the invariance of Haar measure,

$$\begin{aligned} E_{U^* J U} &= \int_{S^{d-1}} \operatorname{sgn}[\xi^* U^* J U \xi] \xi \xi^* d\xi \\ &= \int_{S^{d-1}} \operatorname{sgn}[(U^* \xi)^* U^* J U (U^* \xi)] (U^* \xi)(U^* \xi)^* d\xi \\ &= \int_{S^{d-1}} \operatorname{sgn}[\xi^* U U^* J U U^* \xi] U^* \xi \xi^* U d\xi, \\ &= U^* E_J U. \end{aligned} \quad \blacksquare$$

Recall the definitions of  $\alpha(s, t; a, b)$  and  $\beta(s, t; a, b)$  from Equations (4.4) and (4.5).



**Lemma 5.5.** For  $s, t \in \mathbb{N}$  and  $a, b \geq 0$ ,

$$E_{J(s,t;a,b)} = J(s, t; \alpha(s, t; a, b), \beta(s, t; a, b)) = \alpha(s, t; a, b)I_s \oplus -\beta(s, t; a, b)I_t.$$

Moreover if  $a, b$  are not both 0, then  $\alpha(s, t; a, b)$  and  $\beta(s, t; a, b)$  are not both zero.

*Proof.* If  $X, Y$  are  $s \times s$  and  $t \times t$  orthogonal matrices respectively, then  $U = X \oplus Y \in O(d)$  commutes with  $J := J(s, t; a, b)$  and by Lemma 5.4,

$$U^* E_J U = E_J.$$

It follows that there exists  $\alpha_0$  and  $\beta_0$  such that  $E_J = J(s, t; \alpha_0, \beta_0)$ . That not both  $\alpha_0$  and  $\beta_0$  are zero follows from Lemma 5.2 which says  $E_J \neq 0$ . Finally, in view of Equation (5.3),

$$\alpha_0 = \langle E_J e_1, e_1 \rangle = \int_{S^{d-1}} \operatorname{sgn}[\xi^* J(s, t; a, b) \xi] \xi_1^2 d\xi = \alpha(s, t; a, b)$$

and similarly  $\beta_0 = \beta(s, t; a, b)$ . ■

**Lemma 5.6.** Let  $s, t \in \mathbb{N}_0$  be given and let  $d = s + t$ . Let  $a(s, t), b(s, t)$  denote the pair from Proposition 4.2. Thus,  $a(s, t), b(s, t)$  is uniquely determined by

- (i)  $a(s, t), b(s, t) \geq 0$ ;
- (ii)  $s a(s, t) + t b(s, t) = d$ ;
- (iii)  $\alpha(s, t; a(s, t), b(s, t)) = \beta(s, t; a(s, t), b(s, t))$ ;

and produces the minimum,

- (iv)  $\kappa_*(s, t) = \kappa(s, t; a(s, t), b(s, t)) = d\alpha(s, t; a(s, t), b(s, t))$ .

Abbreviate  $J_* = J(s, t; a(s, t), b(s, t))$ ; then

$$(5.5) \quad E_{J_*} = \frac{\kappa_*(s, t)}{d} J(s, t; 1, 1).$$

*Proof.* Since Proposition 4.2 contains the first four items, it only remains to establish Equation (5.5). From Lemma 5.5 and Item (iii),

$$dE_{J_*} = d\alpha(s, t; a(s, t), b(s, t))J(s, t; 1, 1) = \kappa_*(s, t)J(s, t; 1, 1). \quad \blacksquare$$

**5.3. Dilating to commuting self-adjoint operators.** A  $d \times d$  matrix  $R$  is a **signature matrix** if  $R = R^*$  and  $R^2 = I$ . Thus, a symmetric  $R \in \mathbb{S}_d$  is a signature matrix if its spectrum lies in the set  $\{-1, 1\}$ . Let  $\mathcal{E}(d)$  denote  $d \times d$  signature matrices and  $\mathfrak{C}(d)$  the set of  $d \times d$  symmetric contractions. Thus  $\mathcal{E}(d) \subseteq \mathfrak{C}(d)$ .

**Lemma 5.7.** The set of extreme points of the set of  $\mathfrak{C}(d)$  is  $\mathcal{E}(d)$ . Moreover, each element of  $\mathfrak{C}(d)$  is a (finite) convex combination of elements of  $\mathcal{E}(d)$ .

By Caratheodory's Theorem, there is a bound on the number of summands needed in representing an element of  $\mathfrak{C}(d)$  as a convex combination of elements of  $\mathcal{E}(d)$ .

*Proof.* Suppose for the moment that the set of extreme points of  $\mathfrak{C}(d)$  is a subset of  $\mathcal{E}(d)$ . Since  $\mathcal{E}(d)$  is closed and  $\mathfrak{C}(d)$  is compact, it follows that  $\mathfrak{C}(d) = \overline{\text{co}(\mathcal{E}(d))} = \text{co}(\mathcal{E}(d))$ , where  $\text{co}$  denotes convex hull. Thus the second part of the lemma will follow once it is shown that the extreme points of  $\mathfrak{C}(d)$  lie in  $\mathcal{E}(d)$ .

Suppose  $X \in \mathfrak{C}(d)$  is not in  $\mathcal{E}(d)$ . Without loss of generality, we may assume  $X$  is diagonal with diagonal entries  $\delta_k$ . In particular, there is an  $\ell$  such that  $|\delta_\ell| \neq 1$ . Thus, there exists  $y, z \in (0, 1)$  such that  $y + z = 1$  and  $y - z = \delta_\ell$ . Let  $Y$  denote the diagonal matrix with  $k$ -th diagonal entries  $\delta_k$  for  $j \neq \ell$  and with  $\ell$ -th diagonal entry 1. Define  $Z$  similarly, but with  $\ell$ -th diagonal entry  $-1$ . It follows that  $yY + zZ = X$ . Thus  $X$  is not an extreme point of  $\mathfrak{C}(d)$  and therefore the set of extreme points of  $\mathfrak{C}(d)$  is a subset of  $\mathcal{E}(d)$ .

The proof that each  $E \in \mathcal{E}(d)$  is an extreme point is left to the interested reader.  $\blacksquare$

Our long march of lemmas now culminates in the following lemma. Recall the definitions of  $\kappa_*(s, t)$  and  $\kappa_*(d)$  from the outset of Section 4. A symmetric matrix  $R$  is a **symmetry matrix** if  $R^2$  is projection. Equivalently,  $R$  is a symmetry matrix if  $R = R^*$  and the spectrum of  $R$  lies in the set  $\{-1, 0, 1\}$ . For a symmetric matrix  $D$ , the triple  $\text{sign}(D) = (p, z, n)$ , called the **signature** of  $D$ , denotes the number of positive, zero and negative eigenvalues of  $D$  respectively. Note that a symmetry matrix  $R$  is determined, up to unitary equivalence, by its signature.

**Lemma 5.8.** *If  $R$  is a signature matrix with  $s$  positive eigenvalues and  $t$  negative eigenvalues, then there exists  $D \in \mathfrak{D}(d)$  such that*

$$\kappa_*(s, t)R = V^* M_D V.$$

*In particular, replacing  $D$  with  $D' = \frac{\kappa_*(d)}{\kappa_*(s, t)} D$  and noting that  $\kappa_*(d) \leq \kappa_*(s, t)$ , we have  $D' \in \mathfrak{D}(d)$  and*

$$\kappa_*(d)R = V^* M_{D'} V.$$

*Here  $V$  is the isometry from Lemma 3.4. We emphasize that the  $V$  does not depend on  $R$  or even on  $s, t$ .*

*Proof.* There is an  $s, t$  and unitary  $W$  such that  $R = W^* J(s, t; 1, 1) W$ . From Lemma 5.6, there exists a  $J_* \in \mathbb{S}_d$  such that  $E_{J_*} = \frac{\kappa_*(s, t)}{d} J(s, t; 1, 1)$ . Using Lemmas 3.5, 5.2 and 5.4,

$$\begin{aligned} V^* M_{D_{W^* J_* W}} V &= C_{D_{W^* J_* W}} \\ &= d E_{W^* J_* W} \\ &= W^* d E_{J_*} W \\ &= W^* d \frac{\kappa_*(s, t)}{d} J(s, t; 1, 1) W \\ &= \kappa_*(s, t) W^* J(s, t; 1, 1) W \\ &= \kappa_*(s, t) R. \end{aligned} \quad \blacksquare$$

**Theorem 5.9.** *Given  $d$ , there exists family  $\mathcal{C}_d$  of commuting self-adjoint contractions on a (common) Hilbert space  $\mathcal{H}$  and an isometry  $V: \mathbb{R}^d \rightarrow \mathcal{H}$  such that for each contraction  $C \in \mathbb{S}_d$ , there is a  $T_C \in \mathcal{C}_d$  with*

$$\kappa_*(d)C = V^*T_C V.$$

*Proof.* Set  $\mathcal{H} := L^2(O(d), \mathbb{R}^d)$  and let  $V: \mathbb{R}^d \rightarrow \mathcal{H}$  denote the isometry from Lemma 3.4. Let  $\mathcal{C}_d$  denote the collection of operators  $M_D$  for  $D \in \mathfrak{D}(d)$ . By Lemma 3.3,  $\mathcal{C}_d$  is a collection of commuting operators. By Lemma 3.2, each  $M_D$  is a self-adjoint contraction. Finally, observe that  $\mathcal{C}_d$  is convex.

By Lemma 5.7 there exists an  $h$  and signature matrices  $R_1, \dots, R_h$  such that  $C = \sum_{j=1}^n c_j R_j$ , where  $c_j \geq 0$  and  $\sum c_j = 1$ . By Lemma 5.8, there exists  $S_1, \dots, S_h \in \mathcal{C}_d$  such that  $\kappa_*(d)R_k = V^*S_k V$  for  $k \in \{1, \dots, h\}$ . Hence,  $\kappa_*(d)C = V^*SV$ , where  $S = \sum c_j S_j \in \mathcal{C}_d$ . ■

**5.4. Optimality of  $\kappa_*(d)$ .** The following theorem contains the optimality statement of Theorem 1.1 and Theorem 1.7. It also contains a preliminary version of Theorem 1.6. Recall  $\mathfrak{C}^{(g)}$  is the sequence  $(\mathfrak{C}^{(g)}(d))_d$  and  $\mathfrak{C}^{(g)}(d)$  is the set of  $g$ -tuples of symmetric  $d \times d$  contractions.

**Theorem 5.10.** *For each  $g$  and  $d$ , if  $B$  is any  $g$ -tuple of symmetric  $d \times d$  matrices, then  $[-1, 1]^g \subseteq \mathcal{S}_{L_B}$  implies  $\kappa_*(d)\mathfrak{C}^{(g)} \subseteq \mathcal{D}_{L_B}$ .*

*Conversely, if  $\kappa > \kappa_*(d)$ , then there exists a  $g$  and a  $g$ -tuple of symmetric matrices  $B$  such that  $[-1, 1]^g \subseteq \mathcal{S}_{L_B}$ , but  $\kappa\mathfrak{C}^{(g)} \not\subseteq \mathcal{D}_{L_B}$ .*

*In particular,  $\vartheta(d) = (\kappa_*(d))^{-1}$  is the optimal constant in Theorem 1.1.*

*Proof.* Starting with the proof of the second statement, fix  $d$  and suppose  $\kappa > \kappa_*(d)$ . Let  $(\hat{s}, \hat{t})$  be a pair for which  $\kappa_*(d) = \kappa_*(\hat{s}, \hat{t})$ . Let  $(\hat{a}, \hat{b})$  be a pair of positive numbers such that  $\hat{s}\hat{a} + \hat{t}\hat{b} = d$  and  $\kappa_*(d) = \kappa(\hat{s}, \hat{t}; \hat{a}, \hat{b})$  coming from Lemma 5.6. Let  $\hat{J} = J(\hat{s}, \hat{t}; \hat{a}, \hat{b})$  and define the distinguished (infinite variable pencil)  $\hat{L}: L^\infty(O(d)) \rightarrow \mathbb{S}_d$  by

$$(5.6) \quad \hat{L}(x) = \frac{1}{\kappa_*(d)} \int_{O(d)} U^* \hat{J} U x(U) dU,$$

for  $x \in L^\infty(O(d))$ . By analogy with the sets  $\mathcal{D}_{L_B}$ , let  $\mathcal{D}_{\hat{L}}(n)$  denote those measurable  $X: O(d) \rightarrow \mathbb{S}_n$  such that

$$\hat{L}(X) = \frac{1}{\kappa_*(d)} \int_{O(d)} U^* \hat{J} U \otimes X(U) dU,$$

satisfies  $I - \hat{L}(X) \succeq 0$ .

Let  $\mathfrak{C}^\infty$  denote the sequence of sets  $(\mathfrak{C}^\infty(n))$ , where elements of  $\mathfrak{C}^\infty(n)$  are measurable functions

$$X: O(d) \rightarrow \mathbb{S}_n,$$

such that  $X(U)$  is a symmetric contraction for each  $U \in O(d)$ . In particular,  $x \in \mathfrak{C}^\infty(1)$  is an element of  $L^\infty(O(d))$  of norm at most one and  $\mathfrak{C}^\infty$  can be thought of as an infinite dimensional matrix cube.

Given  $x \in \mathfrak{C}^\infty(1)$  and a unit vector  $e$ , note that

$$e^* \hat{L}(x) e \leq \frac{1}{\kappa_*(d)} \int_{O(d)} |e^* U^* \hat{J} U e| dU = \frac{1}{\kappa_*(d)} \int_{S^{d-1}} |\xi^* \hat{J} \xi| d\xi = 1.$$

Thus  $I - \hat{L}(x) \succeq 0$ . Hence  $\mathfrak{C}^\infty(1) \subseteq \mathcal{D}_{\hat{L}}(1) = \mathcal{S}_{\hat{L}}$ .

Now consider the mapping  $X : O(d) \rightarrow O(d)$  defined by

$$X(U) = U^* J(\hat{s}, \hat{t}; 1, 1) U.$$

In particular,  $X$  pointwise has norm one and thus  $X \in \mathfrak{C}^\infty(d)$ . We next show that  $X \notin \frac{1}{\kappa} \mathcal{D}_{\hat{L}}(d)$ .

For  $U \in O(d)$ , let  $Z(U) = U^* \hat{J} U$ . With  $\mathbf{e} = \frac{1}{\sqrt{d}} \sum_{j=1}^d e_j \otimes e_j$ ,

$$\mathbf{e}^*(Z(U) \otimes X(U)) \mathbf{e} = \frac{1}{d} \sum_{s,t=1}^d e_s^* Z(U) e_t e_s^* X(U) e_t = \frac{1}{d} \langle Z(U), X(U) \rangle_{\text{tr}},$$

where  $\langle \cdot, \cdot \rangle_{\text{tr}}$  is the trace inner product,

$$\langle A, B \rangle_{\text{tr}} = \text{tr}(AB^*) = \sum_{j,k} e_j^* A e_k e_k^* B e_j.$$

Now,

$$\begin{aligned} \text{tr}(Z(U)X(U)) &= \text{tr}(U^* \hat{J} J(\hat{s}, \hat{t}; 1, 1) U) \\ &= \text{tr}(\hat{J} J(\hat{s}, \hat{t}; 1, 1)) = \hat{s}a(\hat{s}, \hat{t}) + \hat{t}b(\hat{s}, \hat{t}) \\ &= d. \end{aligned}$$

Hence

$$\begin{aligned} \mathbf{e}^* \hat{L}(X) \mathbf{e} &= \frac{1}{\kappa_*(d)} \int \mathbf{e}^*(Z(U) \otimes X(U)) \mathbf{e} dU \\ &= \frac{1}{\kappa_*(d)} \frac{1}{d} d. \end{aligned}$$

Thus  $\|\hat{L}(X)\| \geq \frac{1}{\kappa_*(d)} > \frac{1}{\kappa}$ , so

$$\frac{1}{\kappa} I - \hat{L}(X) \not\succeq 0$$

as predicted.

We next realize  $\hat{L}$  as a limit of pencils  $L_B$  with  $B \in \mathbb{S}_d^g$ . Suppose  $(\mathcal{P}_k)$  is a sequence of (measurable) partitions of  $O(d)$  and write  $\mathcal{P}_k = \{P_{k,1}, \dots, P_{k,g_k}\}$ . Consider the corresponding  $g_k$ -tuples  $A^k = (A_1^k, \dots, A_{g_k}^k) \in \mathbb{S}_d^{g_k}$ , where

$$A_j^k = \frac{1}{\kappa_*(d)} \int_{P_{k,j}} U^* \hat{J} U dU = \int_{P_{k,j}} Z(U) dU,$$

and the associated homogeneous linear pencil,

$$L_k(x) = \sum_{j=1}^{g_k} A_j^k x_j.$$

Given  $y \in \mathfrak{C}^{(g_k)}(1) = [-1, 1]^{g_k}$ , let  $x = \sum_{j=1}^{g_k} y_j \chi_{P_{k,j}}$ , where  $\chi_P$  denotes the characteristic function of the set  $P$ . Since  $x \in \mathfrak{C}^\infty(1)$  and

$$L_k(y) = \hat{L}(x),$$

it follows that  $y \in \mathcal{S}_{L_{A^k}}$ . Thus,  $[-1, 1]^{g_k} \subseteq \mathcal{S}_{L_{A^k}}$ .

Suppose that  $U_{k,j}$  are given points in  $P_{k,j}$ . Given  $k$ , let  $X^k = (X_1^k, \dots, X_{g_k}^k)$  where

$$X_j^k = X(U_{k,j}) = U_{k,j}^* J(\hat{s}, \hat{t}; 1, 1) U_{k,j}.$$

In particular,  $\|X_j^k\| \leq 1$ . Evaluate,

$$L_k(X^k) = \frac{1}{\kappa_*(d)} \sum_{j=1}^{g_k} A_{k,j} \otimes X_{k,j}.$$

Hence

$$\hat{L}(X) - L_k(X^k) = \sum_{j=1}^{g_k} \int_{P_{k,j}} Z(U) \otimes (X(U) - X(U_{k,j})) dU.$$

The uniform continuity of  $X(U)$  implies there exists a choice of  $P_{k,j}$  and  $U_{k,j}$  such that  $L_k(X^k)$  converges to  $\hat{L}(X)$ . Hence,  $\|L_k(X^k)\| > \kappa$  for sufficiently large  $k$ . Consequently  $X \in \mathfrak{C}^{g_k}(d)$ , but  $\kappa X \notin \mathcal{D}_{L_{A^k}}(d)$  and the proof of the second statement is complete.

Turning to the first part of the theorem, suppose that  $B$  is a  $g$ -tuple of  $d \times d$  symmetric matrices and  $[-1, 1]^g \subseteq \mathcal{S}_{L_B}$ . Given a  $g$ -tuple  $X \in \mathfrak{C}^{(g)}(d)$ , Theorem 5.9 produces a Hilbert space  $\mathcal{H}$ , a  $g$ -tuple of commuting self-adjoint contractions on  $\mathcal{H}$ , an isometry  $V : \mathbb{R}^d \rightarrow \mathcal{H}$  such that

$$\kappa_*(d)X_j = V^* T_j V,$$

a relationship summarized by  $\kappa_*(d)X = V^* T V$ . By Proposition 2.1,  $\kappa_*(d)\mathfrak{C}^{(g)} \subseteq \mathcal{D}_{L_B}$  and the proof of the first statement of the theorem is complete.

For the last statement in the theorem, suppose  $\kappa$  has the property that for every  $g$  each  $g$ -tuple of commuting symmetric matrices of size  $d$  dilates to a tuple of commuting symmetric contractions on Hilbert space. Proposition 2.1 implies  $\kappa\mathfrak{C}^{(g)} \subseteq \mathcal{D}_{L_B}$  for any  $g$ -tuple  $B$  of symmetric matrices of size  $d$  such that  $[-1, 1]^g \subseteq \mathcal{S}_{L_B}$ . Hence, by what has already been proved,  $\kappa \leq \kappa_*(d)$ . ■

## 6. THE OPTIMALITY CONDITION $\alpha = \beta$ IN TERMS OF BETA FUNCTIONS

In this section  $\alpha(s, t; a, b)$  and  $\beta(s, t; a, b)$  which were defined in Equations (4.4) and (4.5) (see also Lemma 5.5) are computed in terms of the regularized incomplete beta function. See Lemma 6.6. A consequence is the relation of Equation (1.5). Lemma 6.5 figures in the proof of Theorem 1.6 in Section 7.

Let  $\Gamma$  denote the Euler gamma function [Rai71].

**Lemma 6.1.** *Suppose  $m \in \mathbb{R}_{\geq 0}$ . Then*

$$\int_0^\infty r^m e^{-r^2} dr = \frac{1}{2} \Gamma\left(\frac{m+1}{2}\right).$$

*Proof.* Setting  $s := r^2$ , we have  $r^m = s^{\frac{m}{2}}$  and  $\frac{ds}{dr} = \frac{dr^2}{dr} = 2r$ , i.e.,  $dr = \frac{ds}{2r} = \frac{ds}{2\sqrt{s}}$ . Then

$$\int_0^\infty r^m e^{-r^2} dr = \int_0^\infty \frac{s^{\frac{m}{2}}}{2\sqrt{s}} ds = \frac{1}{2} \int_0^\infty s^{\frac{m-1}{2}} ds = \frac{1}{2} \Gamma\left(\frac{m+1}{2}\right). \quad \blacksquare$$

**Lemma 6.2.**

$$\int_{\mathbb{R}^n} e^{-\|x\|^2} dx = \pi^{\frac{n}{2}}.$$

*Proof.*

$$\int_{\mathbb{R}^n} e^{-\|x\|^2} dx = \int_{\mathbb{R}} \dots \int_{\mathbb{R}} e^{-x_1^2} \dots e^{-x_n^2} dx_n \dots dx_1 = \left( \int_{\mathbb{R}} e^{-x^2} dx \right)^n \stackrel{6.1}{=} \Gamma\left(\frac{1}{2}\right)^n = \pi^{\frac{n}{2}}. \quad \blacksquare$$

We equip the unit sphere in  $S^{n-1} \subseteq \mathbb{R}^n$  with the unique rotation invariant probability measure.

**Remark 6.3.** Recall that the surface area of the  $n-1$ -dimensional unit sphere  $S^{n-1} \subseteq \mathbb{R}^n$  is

$$\text{area}(S^{n-1}) = \frac{n\pi^{\frac{n}{2}}}{\Gamma(1 + \frac{n}{2})} = \frac{2\pi^{\frac{n}{2}}}{\Gamma(\frac{n}{2})}.$$

□

Now we come to a key step, converting integrals over the sphere  $S^{d-1}$  to integrals over  $\mathbb{R}^d$ .

**Lemma 6.4.** Suppose  $A \in \mathbb{R}^{d \times d}$  and  $f: \mathbb{R}^d \rightarrow \mathbb{R}$  is quadratically homogeneous, i.e.,  $f(\lambda x) = \lambda^2 f(x)$  for all  $x \in \mathbb{R}^d$  and  $\lambda \in \mathbb{R}$ . Suppose furthermore that  $f|_{S^{d-1}}$  is integrable on  $S^{d-1}$ . Then

$$\begin{aligned} \int_{S^{d-1}} f(\xi) d\xi &= \frac{2}{d\pi^{\frac{d}{2}}} \int_{\mathbb{R}^d} f(x) e^{-\|x\|^2} dx \quad \text{and} \\ d \int_{S^{d-1}} f(\xi) d\xi &= \frac{1}{(2\pi)^{\frac{d}{2}}} \int_{\mathbb{R}^d} f(x) e^{-\frac{\|x\|^2}{2}} dx. \end{aligned}$$

*Proof.* The first equality follows from

$$\begin{aligned} \int_{\mathbb{R}^d} f(x) e^{-\|x\|^2} dx &= \int_{S^{d-1}} \int_0^\infty \text{area}(rS^{d-1}) f(r\xi) e^{-\|r\xi\|^2} dr d\xi \\ &= \int_{S^{d-1}} \int_0^\infty r^{d-1} \text{area}(S^{d-1}) r^2 f(\xi) e^{-r^2} dr d\xi \\ &= \text{area}(S^{d-1}) \left( \int_0^\infty r^{d+1} e^{-r^2} dr \right) \int_{S^{d-1}} f(\xi) d\xi \\ &= \frac{d\pi^{\frac{d}{2}}}{\Gamma(1 + \frac{d}{2})} \frac{1}{2} \Gamma\left(1 + \frac{d}{2}\right) \int_{S^{d-1}} f(\xi) d\xi. \end{aligned}$$

where the last equation uses Remark 6.3 and Lemma 6.1. The proof of the second equality is similar and is left as an exercise for the reader. ■

**Lemma 6.5.** *Suppose  $J \in \mathbb{R}^{d \times d}$  is any matrix and consider the zero matrix  $0_u := 0 \in \mathbb{R}^{u \times u}$ . Suppose also  $i, j \in \{1, \dots, d\}$ . Then there is some  $\gamma \in \mathbb{R}$  such that*

$$\int_{S^{d+u-1}} \operatorname{sgn} [\xi^* (J \oplus 0_u) \xi] \xi_i \xi_j d\xi = \begin{cases} \frac{d}{d+u} \int_{S^{d-1}} \operatorname{sgn} [\xi^* J \xi] \xi_i \xi_j d\xi & \text{if } i, j \in \{1, \dots, d\} \\ \gamma & \text{if } i = j \in \{d+1, \dots, d+u\} \\ 0 & \text{otherwise} \end{cases}$$

*Proof.* Set

$$C := \frac{1}{2}(d+u)\pi^{\frac{d+u}{2}} \quad \text{and} \quad c := \frac{1}{2}d\pi^{\frac{d}{2}}.$$

By Lemma 6.4, the left hand side equals

$$\frac{1}{C} \int_{\mathbb{R}^{d+u}} \operatorname{sgn}[x^*(J \oplus 0_s)x] x_i x_j e^{-\|x\|^2} dx.$$

If  $i, j \in \{1, \dots, d\}$ , then this in turn equals

$$\begin{aligned} \frac{1}{C} \int_{\mathbb{R}^d} \int_{\mathbb{R}^u} \operatorname{sgn}[y^* J y] y_i y_j e^{-\|y\|^2 - \|z\|^2} dz dy &= \frac{1}{C} \left( \int_{\mathbb{R}^u} e^{-\|z\|^2} dz \right) \int_{\mathbb{R}^d} \operatorname{sgn}[y^* J y] y_i y_j e^{-\|y\|^2} dy \\ &= \frac{c}{C} \pi^{\frac{u}{2}} \int_{S^{d-1}} \operatorname{sgn}[\xi^* J \xi] \xi_i \xi_j d\xi \end{aligned}$$

where the last equality follows from Lemmas 6.2 and 6.4. If  $i, j \in \{d+1, \dots, d+u\}$ , then it equals

$$\frac{1}{C} \int_{\mathbb{R}^d} \int_{\mathbb{R}^u} \operatorname{sgn}[y^* J y] z_{i-d} z_{j-d} e^{-\|y\|^2 - \|z\|^2} dz dy$$

which equals up to a constant depending only on  $J$  the integral

$$\int_{\mathbb{R}^u} z_{i-d} z_{j-d} e^{-\|z\|^2} dz$$

which is zero for symmetry reasons if  $i \neq j$  and which depends only on  $u$  if  $i = j$ . The remaining case where one of  $i$  and  $j$  is in  $\{1, \dots, d\}$  and the other one in  $\{d+1, \dots, d+u\}$  follows similarly.  $\blacksquare$

**Lemma 6.6.** *Let  $s, t \in \mathbb{N}$ ,  $d := s + t$  and  $a, b \in \mathbb{R}_{\geq 0}$  with  $a + b > 0$ . Then*

$$(6.1) \quad \alpha(s, t; a, b) := \int_{S^{d-1}} \operatorname{sgn}[\xi^* J(s, t; a, b) \xi] \xi_i^2 d\xi = \frac{1}{d} \left( 2I_{\frac{a}{a+b}} \left( \frac{t}{2}, \frac{s}{2} + 1 \right) - 1 \right)$$

for all  $i \in \{1, \dots, s\}$ . Analogously,

$$(6.2) \quad \beta(s, t; a, b) := - \int_{S^{d-1}} \operatorname{sgn}[\xi^* J(s, t; a, b) \xi] \xi_i^2 d\xi = \frac{1}{d} \left( 2I_{\frac{b}{a+b}} \left( \frac{s}{2}, \frac{t}{2} + 1 \right) - 1 \right)$$

for all  $i \in \{s+1, \dots, s+t\}$ . An additional obvious property is

$$(6.3) \quad \alpha(s, t; a, b) = \beta(t, s; b, a).$$

*Proof.* We have

$$\begin{aligned}
& \int_{S^{d-1}} \operatorname{sgn}[\xi^* J(s, t; a, b) \xi] \xi_i^2 d\xi \\
& \stackrel{6.4}{=} \frac{2}{d\pi^{\frac{d}{2}}} \int_{\mathbb{R}^d} \operatorname{sgn}[x^* J(s, t; a, b) x] x_i^2 e^{-\|x\|^2} dx \\
& = \frac{2}{sd\pi^{\frac{d}{2}}} \int_{\mathbb{R}^d} \operatorname{sgn}[x^* J(s, t; a, b) x] (x_1^2 + \cdots + x_s^2) e^{-\|x\|^2} dx \\
& = \frac{2}{sd\pi^{\frac{d}{2}}} \int_{\mathbb{R}^s} \int_{\mathbb{R}^t} \operatorname{sgn}[a\|y\|^2 - b\|z\|^2] \|y\|^2 e^{-\|y\|^2 - \|z\|^2} dz dy \\
& = \frac{2}{sd\pi^{\frac{d}{2}}} \int_0^\infty \operatorname{area}(\sigma S^{s-1}) \int_0^\infty \operatorname{area}(\tau S^{t-1}) \operatorname{sgn}[a\sigma^2 - b\tau^2] \sigma^2 e^{-\sigma^2 - \tau^2} d\tau d\sigma \\
& = \frac{2}{sd\pi^{\frac{d}{2}}} \int_0^\infty \sigma^{s-1} \frac{2\pi^{\frac{s}{2}}}{\Gamma(\frac{s}{2})} \int_0^\infty \tau^{t-1} \frac{2\pi^{\frac{t}{2}}}{\Gamma(\frac{t}{2})} \operatorname{sgn}[a\sigma^2 - b\tau^2] \sigma^2 e^{-\sigma^2 - \tau^2} d\tau d\sigma \\
& = \frac{8}{sd\Gamma(\frac{s}{2})\Gamma(\frac{t}{2})} \int_0^\infty \int_0^\infty \sigma^{s+1} \tau^{t-1} \operatorname{sgn}[a\sigma^2 - b\tau^2] e^{-\sigma^2 - \tau^2} d\tau d\sigma \\
& = \frac{8}{sd\Gamma(\frac{s}{2})\Gamma(\frac{t}{2})} \int_0^\infty r \int_0^{\frac{\pi}{2}} (r \cos \varphi)^{s+1} (r \sin \varphi)^{t-1} \operatorname{sgn}[a(\cos \varphi)^2 - b(\sin \varphi)^2] e^{-r^2} d\varphi dr \\
& = \frac{8}{sd\Gamma(\frac{s}{2})\Gamma(\frac{t}{2})} \left( \int_0^\infty r^{d+1} e^{-r^2} dr \right) \int_0^{\frac{\pi}{2}} (\cos \varphi)^{s+1} (\sin \varphi)^{t-1} \operatorname{sgn}[a(\cos \varphi)^2 - b(\sin \varphi)^2] d\varphi \\
& \stackrel{6.1}{=} \frac{4\Gamma(\frac{d}{2} + 1)}{sd\Gamma(\frac{s}{2})\Gamma(\frac{t}{2})} \int_0^{\frac{\pi}{2}} (\cos \varphi)^{s+1} (\sin \varphi)^{t-1} \operatorname{sgn}[a(\cos \varphi)^2 - b(\sin \varphi)^2] d\varphi \\
& = \frac{\Gamma(\frac{d}{2})}{s\Gamma(\frac{s}{2})\Gamma(\frac{t}{2})} \int_0^1 (1-x)^{\frac{s+1-1}{2}} x^{\frac{t-1-1}{2}} \operatorname{sgn}[a(1-x) - bx] dx \\
& = \frac{1}{sB(\frac{s}{2}, \frac{t}{2})} \int_0^1 (1-x)^{\frac{s}{2}} x^{\frac{t}{2}-1} \operatorname{sgn}[a - (a+b)x] dx
\end{aligned}$$

using a change of variable  $x = (\sin \varphi)^2$  which makes

$$\frac{dx}{d\varphi} = 2(\sin \varphi)(\cos \varphi) = 2\sqrt{x}\sqrt{1-x}.$$

Now suppose that  $a, b \in \mathbb{R}_{\geq 0}$  with  $a + b > 0$ . Then the integral in the last expression equals

$$\begin{aligned}
& \int_0^{\frac{a}{a+b}} (1-x)^{\frac{s}{2}} x^{\frac{t}{2}-1} dx - \int_{\frac{a}{a+b}}^1 (1-x)^{\frac{s}{2}} x^{\frac{t}{2}-1} dx \\
& = B_{\frac{a}{a+b}}\left(\frac{t}{2}, \frac{s}{2} + 1\right) - \int_0^{\frac{b}{a+b}} x^{\frac{s}{2}} (1-x)^{\frac{t}{2}-1} dx \\
& = B_{\frac{a}{a+b}}\left(\frac{t}{2}, \frac{s}{2} + 1\right) - B_{\frac{b}{a+b}}\left(\frac{s}{2} + 1, \frac{t}{2}\right).
\end{aligned}$$



Using

$$B\left(\frac{s}{2}, \frac{t}{2}\right) = \frac{\Gamma\left(\frac{s}{2}\right)\Gamma\left(\frac{t}{2}\right)}{\Gamma\left(\frac{d}{2}\right)} = \frac{d}{s} \frac{\frac{s}{2}\Gamma\left(\frac{s}{2}\right)\Gamma\left(\frac{t}{2}\right)}{\frac{d}{2}\Gamma\left(\frac{d}{2}\right)} = \frac{d}{s} \frac{\Gamma\left(\frac{s}{2}+1\right)\Gamma\left(\frac{t}{2}\right)}{\frac{d}{2}\Gamma\left(\frac{d}{2}+1\right)} = \frac{d}{s} B\left(\frac{s}{2}+1, \frac{t}{2}\right),$$

we see that

$$\alpha(s, t; a, b) = \frac{1}{d} \left( I_{\frac{a}{a+b}}\left(\frac{t}{2}, \frac{s}{2}+1\right) - I_{\frac{b}{a+b}}\left(\frac{s}{2}+1, \frac{t}{2}\right) \right)$$

where  $I$  denotes the regularized (incomplete) beta function. Finally, (6.1) follows using

$$I_{1-p}(\zeta, \eta) = 1 - I_p(\eta, \zeta).$$

The proof of (6.2) is similar. ■

## 7. RANK VERSUS SIZE FOR THE MATRIX CUBE

In this section we show how to pass from size  $d$  to rank  $d$  in the first part of Theorem 5.10, thus completing our dilation theoretic proof of Theorem 1.6. Accordingly fix, for the remainder of this section, positive integers  $d \leq m$ .

Given positive integers  $s, t, u$  and numbers  $a, b, c$ , let

$$J(s, t, u; a, b, c) = aI_s \oplus -bI_t \oplus cI_u.$$

**Lemma 7.1.** *Given positive integers  $s, t$  with  $s + t = d$  and nonnegative numbers  $a, b, c$ , there exists real numbers  $\alpha, \beta$ , and  $\gamma$  such that*

$$J(s, t, m-d; \alpha, \beta, \gamma) = m \int_{S^{m-1}} \text{sgn}[\xi^* J(s, t, m-d; a, b, c) \xi] \xi \xi^* d\xi.$$

*Proof.* Given  $U_v \in \mathcal{O}(v)$ , for  $v = s, t, m-d$ , let  $U$  denote the block diagonal matrix with entries  $U_s, U_t, U_{m-d}$ . Thus,  $U \in \mathcal{O}(m)$  and  $U$  commutes with  $J(s, t, m-d; a, b, c)$ . The conclusion now follows, just as in Lemma 5.5. ■

**Lemma 7.2.** *For each  $s, t$  with  $s + t = d$ , there exists a  $\gamma = \gamma(s, t)$  such that*

$$(7.1) \quad \kappa_*(s, t) J(s, t, u; 1, 1, \gamma(s, t)) = m \int_{S^{m-1}} \text{sgn}[\xi^* J(s, t, m-d; a(s, t), b(s, t), 0) \xi] \xi \xi^* d\xi.$$

Here  $\kappa_*(s, t), a(s, t)$  and  $b(s, t)$  are the optimal choices from Proposition 4.2.

*Proof.* Denote the right hand side of (7.1) by  $E$ . Then by Lemma 6.5,

$$e_i E e_j = \begin{cases} d \int_{S^{d-1}} \text{sgn}[\xi^* J(s, t; a(s, t), b(s, t)) \xi] e_i \xi \xi^* e_j d\xi & \text{if } i = j \in \{1, \dots, d\} \\ \gamma & \text{if } i = j \in \{d+1, \dots, m\} \\ 0 & \text{otherwise} \end{cases}$$

for some  $\gamma \in \mathbb{R}$  and all  $i, j \in \{1, \dots, m\}$ . On the other hand, from Lemma 5.6,

$$\frac{\kappa_*(s, t)}{d} J(s, t; 1, 1) = \int_{S^{d-1}} \text{sgn}[\xi^* J(s, t; a(s, t), b(s, t)) \xi] \xi \xi^* d\xi.$$

Hence, with  $P$  denoting the projection of  $\mathbb{R}^d \oplus \mathbb{R}^{m-d}$  onto the first  $d$  coordinates,

$$PEP = \kappa_*(s, t)J(s, t; 1, 1)$$

and the conclusion of the lemma follows.  $\blacksquare$

Let  $\mathcal{H}$  denote the Hilbert space  $\mathbb{R}^m \otimes L^2(O(m))$  and let  $V : \mathbb{R}^m \rightarrow \mathcal{H}$  denote the isometry,

$$Vx(U) = x.$$

Thus  $\mathcal{H}$  and  $V$  are the Hilbert space and isometry (with  $m$  in place of  $d$ ) from Equations (3.2) and (3.3). Recall too the collection  $\mathfrak{D}(m)$  of contractive measurable mappings  $D : O(m) \rightarrow M_m$  taking diagonal values, and, for  $D \in \mathfrak{D}(m)$ , the contraction operator  $M_D : \mathcal{H} \rightarrow \mathcal{H}$ .

**Lemma 7.3.** *For each  $m \times m$  symmetry matrix  $R$  of rank  $d$  there exists a  $D \in \mathfrak{D}(m)$  such that*

$$\kappa_*(d) PRP = PV^*M_DVP,$$

where  $P$  is the projection onto the range of  $R$ .

*Proof.* The proof is similar to the proof of Lemma 5.8. Let  $s$  and  $t$  denote the number of positive and negative eigenvalues of  $R$ . Hence,  $R = W^*J(s, t, m-d; 1, 1, 0)W$  for some  $m \times m$  unitary  $W$ . Let  $J_* = J(s, t, m-d; a(s, t), b(s, t), 0)$  and define  $D \in O(m)$  by

$$D(U) = \sum_{j=1}^m \operatorname{sgn}[e_j^* U^* W^* J W U e_j] e_j e_j^* dU.$$

Now, by Lemma 3.5, Remark 5.3 and Lemma 7.2,

$$\begin{aligned} V^* M_{D_{W^* J_* W}} V &= C_{D_{W^* J_* W}} \\ &= \int_{O(m)} U D(U) U^* dU \\ &= \sum_{j=1}^m \int_{O(m)} \operatorname{sgn}[e_j^* U^* W^* J W U e_j] U e_j e_j^* U^* dU \\ &= \sum_{j=1}^m \int_{O(m)} \operatorname{sgn}[e_j^* U^* J U e_j] W^* U e_j e_j^* U^* W dU \\ &= W^* \left( \sum_{j=1}^m \int_{O(m)} \operatorname{sgn}[e_j^* U^* J U e_j] U e_j e_j^* U^* dU \right) W \\ &= m W^* \left( \int_{S^{m-1}} \operatorname{sgn}[\xi^* J \xi] \xi \xi^* d\xi \right) W \\ &= \kappa_*(s, t) W^* J(s, t, m-d; 1, 1, \gamma(s, t)) W. \end{aligned}$$

The observation

$$PW^*J(s, t, m-d; 1, 1, \gamma(s, t))WP = PW^*J(s, t, m-d; 1, 1, 0)WP$$

completes the proof.  $\blacksquare$

Given a  $g$ -tuple  $\mathcal{M} = (\mathcal{M}_1, \dots, \mathcal{M}_g)$  of  $d$ -dimensional subspaces of  $\mathbb{R}^m$ , let  $\mathfrak{C}(\mathcal{M})$  denote the collection of  $g$ -tuples of  $m \times m$  symmetric contractions  $C = (C_1, \dots, C_g)$  where each  $C_j$  has range in  $\mathcal{M}_j$ .

**Lemma 7.4.** *The set  $\mathfrak{C}(\mathcal{M})$  is closed and convex and its extreme points are the tuples of the form  $E = (E_1, \dots, E_g)$ , where each  $E_j$  is a symmetry matrix with rank  $d$ .*

*Proof.* Given a subspace  $\mathcal{N}$  of  $\mathbb{R}^m$  of dimension  $d$ , note that the set  $n \times n$  symmetric contractions with range in  $\mathcal{N}$  is a convex set whose extreme points are symmetry matrices whose range is exactly  $\mathcal{N}$  (cf. Lemma 5.7). Since  $\mathcal{M} = \times_{j=1}^g \mathcal{M}_j$  the result follows.  $\blacksquare$

**Lemma 7.5.** *Suppose  $X = (X_1, \dots, X_g)$  is a tuple of  $m \times m$  symmetric contractions. If  $P_1, \dots, P_g$  is a tuple of rank  $d$  projections, then there exists a tuple of  $m \times m$  symmetric contractions  $Y = (Y_1, \dots, Y_g)$  such that*

- (i)  $P_j X_j P_j = P_j Y_j P_j$ ; and
- (ii) *there exists a tuple of commuting self-adjoint contractions  $Z = (Z_1, \dots, Z_g)$  on a Hilbert space  $\mathcal{H}$  such that  $\kappa_*(d)Y$  lifts to  $Z$ .*

Thus, there exists an isometry  $Q : \mathbb{R}^m \rightarrow \mathcal{H}$  such that

$$Y_j = \frac{1}{\kappa_*(d)} W^* Z_j W \quad \text{and} \quad P_j W^* Z_j W P_j = P_j X_j P_j.$$

*Proof.* Let  $\mathcal{M}_j$  denote the range of  $P_j$ . Let  $C_j = P_j X_j P_j$ . By Lemma 7.4, there exists a positive integer  $N$  and extreme points  $E^1, \dots, E^N$  in  $\mathfrak{C}(\mathcal{M})$  and positive numbers  $\epsilon_1, \dots, \epsilon_N$  such that

$$C_j = \sum_{k=1}^N \epsilon_k E_j^k.$$

For each  $k, j$  there exist positive integers  $s_j^k, t_j^k$  such that  $s_j^k + t_j^k = 1$  and a unitary matrix  $W_j^k$  such that  $(W_j^k)^* \mathbb{R}^d \oplus \{0\} = \mathcal{M}_j$  and

$$E_j^k = (W_j^k)^* J(s_j^k, t_j^k, m-d, 1, 1, 0) W_j^k.$$

In particular,

$$E_j^k = P_j E_j^k P_j.$$

By Lemma 7.3, there exists  $D_j^k \in \mathfrak{D}(m)$  such that

$$\kappa_*(d) E_j^k = P_j V^* M_{D_j^k} V P_j.$$

Let

$$Z_j = \sum_k M_{D_j^k}.$$

Thus  $Z$  is a  $g$ -tuple of commuting contractions and

$$\begin{aligned} P_j V^* Z_j W P_j &= \sum_{k=1}^N \epsilon_k P_j V^* \mathcal{M}_{D_j^k} V P_j \\ &= \sum_{k=1}^N \epsilon_k E_j^k \\ &= C_j. \end{aligned}$$

Choosing  $Y_j = V^* Z_j V$  completes the proof since the  $Z_j$  are commuting self-adjoint contractions (and  $V$  is an isometry independent of  $j$ ).  $\blacksquare$

**7.1. Proof of Theorem 1.6.** Our dilation theoretic proof of Theorem 1.6 concludes in this subsection. Accordingly, suppose  $B = (B_1, \dots, B_g)$  is a given  $g$ -tuple of  $m \times m$  symmetric matrices of rank at most  $d$  and  $[-1, 1]^g \subseteq \mathcal{S}_B$ . We are to show  $\kappa_*(d)\mathfrak{C}^{(g)} \subseteq \mathcal{D}_B$ .

Let

$$\Lambda_B(x) = \sum_{j=1}^g B_j x_j$$

be the homogeneous linear pencil associated with  $B$ . The aim is to show that  $\Lambda_B(\kappa_*(d)X) \preceq I$  for tuples  $X \in \mathfrak{C}^{(g)}$  and, by Lemma 2.3, it suffices to suppose  $X$  has size  $m$ . Let  $x \in \mathbb{R}^m \otimes \mathbb{R}^m$  be a given unit vector. The proof reduces to showing

$$\kappa_*(d) \langle \Lambda_B(X)x, x \rangle \leq 1.$$

Fix  $j$  and let  $\{f_{j1}, f_{j2}, \dots, f_{jd}\}$  denote an orthonormal basis for the range of  $B_j$  (or any  $d$ -dimensional subspace that contains the range of  $B_j$ ). This uses the rank at most  $d$  assumption. Extend this basis to an orthonormal basis  $\{f_{j1}, \dots, f_{jm}\}$  of all  $\mathbb{R}^m$ . Note that  $f_{jp} \in \{f_{j1}, f_{j2}, \dots, f_{jd}\}^\perp \subseteq (\text{im } B_j)^\perp = \ker B_j$  for all  $j \in \{1, \dots, g\}$  and  $p \in \{d+1, \dots, m\}$  since  $B_j$  is symmetric. The unit vector  $x$  can be written in  $g$  different ways (indexed by  $j \in \{1, \dots, g\}$ ) as

$$x = \sum_{p=1}^m f_{jp} \otimes x_{jp},$$

for vectors  $x_{jp} \in \mathbb{R}^m$ . Let  $P_j$  be the orthogonal projection onto

$$\mathcal{M}_j := \text{span}(\{x_{j1}, \dots, x_{jd}\})$$

and compute for  $j$  fixed and any  $m \times m$  tuple  $Y$  such that  $P_j Y_j P_j = P_j X_j P_j$ ,

$$\begin{aligned}
\langle (B_j \otimes X_j)x, x \rangle &= \sum_{p,q=1}^m \langle B_j f_{jp}, f_{jq} \rangle \langle X_j x_{jp}, x_{jq} \rangle \\
&= \sum_{p,q=1}^d \langle B_j f_{jp}, f_{jq} \rangle \langle X_j x_{jp}, x_{jq} \rangle \\
&= \sum_{p,q=1}^d \langle B_j f_{jp}, f_{jq} \rangle \langle P_j X_j P_j x_{jp}, x_{jq} \rangle \\
&= \sum_{p,q=1}^d \langle B_j f_{jp}, f_{jq} \rangle \langle P_j Y_j P_j x_{jp}, x_{jq} \rangle \\
&= \sum_{p,q=1}^d \langle B_j f_{jp}, f_{jq} \rangle \langle Y_j x_{jp}, x_{jq} \rangle \\
&= \sum_{p,q=1}^m \langle B_j f_{jp}, f_{jq} \rangle \langle Y_j x_{jp}, x_{jq} \rangle \\
&= \langle (B_j \otimes Y_j)x, x \rangle.
\end{aligned}$$

From Lemma 7.5 there exists a Hilbert space  $\mathcal{K}$  (infinite dimensional generally), an isometry  $V: \mathbb{R}^m \rightarrow \mathcal{K}$  and a tuple of commuting self-adjoint contractions  $Z = (Z_1, \dots, Z_g)$  acting on  $\mathcal{K}$  such that  $Y_j$ , defined by  $\kappa_*(d)Y_j = V^* Z_j V$ , satisfies  $P_j Y_j P_j = P_j X_j P_j$ . Hence,

$$\begin{aligned}
\kappa_*(d)\langle \Lambda_B(X)x, x \rangle &= \kappa_*(d)\langle \Lambda_B(Y)x, x \rangle \\
&= \langle (I_m \otimes V^*)\Lambda_B(Z)(I_m \otimes V)x, x \rangle \\
&= \langle \Lambda_B(Z)z, z \rangle,
\end{aligned}$$

where  $z = (I \otimes V)x$ . In particular,  $z$  is a unit vector. Since  $Z$  is a commuting tuple of self-adjoint contractions, just as in Proposition 2.1, the inclusion  $[-1, 1]^g \subseteq \mathcal{S}_{L_B}$  implies,

$$\langle \Lambda_B(Z)z, z \rangle \leq 1.$$

The final conclusion is

$$\kappa_*(d)\langle \Lambda_B(X)x, x \rangle \leq 1. \quad \blacksquare$$

## 8. FREE SPECTRAHEDRAL INCLUSION GENERALITIES

This section begins with a bound on the inclusion scale which depends little on the LMIs involved, Section 8.1. In Subsection 8.2 we prove that the inclusion scale equals the commutability index, that is, Theorem 1.4. In summary, all the claims made in Section 1.3 are established here.

**8.1. A general bound on the inclusion scale.** This subsection gives a bound on the inclusion scale which depends little on the LMIs involved. Recall  $\mathcal{S}_{L_A}$  is the spectrahedron  $\mathcal{D}_{L_A}(1)$  determined by the tuple  $A$ .

**Proposition 8.1.** *Suppose  $A$  and  $B$  are  $g$ -tuples of symmetric matrices, where the  $B_j$  are  $d \times d$  matrices. Suppose further that  $-\mathcal{D}_{L_A} \subseteq \mathcal{D}_{L_A}$ . If  $\mathcal{S}_{L_A} \subseteq \mathcal{S}_{L_B}$ , then  $\mathcal{D}_{L_A}(n) \subseteq d \mathcal{D}_{L_B}(n)$  for each  $n$ .*

**Lemma 8.2.** *Suppose  $T = (T_{j,\ell})$  is a  $d \times d$  block matrix with blocks of equal square size. If  $\|T_{j,\ell}\| \leq 1$  for every  $j, \ell$ , then  $\|T\| \leq d$ .*

*Proof.* Recall that the Cauchy-Schwarz inequality applied with one of the vectors being the all ones vector gives the relation between the 1-norm and 2-norm on  $\mathbb{R}^d$ , namely

$$\left(\sum_{j=1}^d a_j\right)^2 \leq d \sum_{j=1}^d a_j^2 \quad \text{for all } a_1, \dots, a_d \in \mathbb{R}.$$

Consider a vector  $x = \sum_{\ell=1}^d x_\ell \otimes e_\ell$  and estimate,

$$\begin{aligned} \|Tx\|^2 &= \sum_{j=1}^d \left\| \sum_{\ell=1}^d T_{j,\ell} x_\ell \right\|^2 \\ &\leq \sum_{j=1}^d \left( \sum_{\ell=1}^d \|x_\ell\| \right)^2 \\ &\leq \sum_{j=1}^d d \sum_{\ell=1}^d \|x_\ell\|^2 \\ &= \sum_{j=1}^d d \|x\|^2 \\ &= d^2 \|x\|^2. \end{aligned}$$

Thus,  $\|Tx\| \leq d\|x\|$ . ■

*Proof of Proposition 8.1.* Let  $\{e_s\}$  denote the standard orthonormal basis for  $\mathbb{R}^n$ . Fix  $1 \leq s \neq t \leq n$  and set  $p_{s,t}^\pm := \frac{1}{\sqrt{2}}(e_s \pm e_t) \in \mathbb{R}^n$ . In particular, with

$$P_{s,t}^\pm = I_d \otimes p_{s,t}^\pm,$$

the orthonormality of the basis gives,

$$(P_{s,t}^\pm)^* P_{s,t}^\pm = I_d.$$

Moreover, for  $d \times d$  matrix  $C$  and  $n \times n$  matrix  $M$ ,

$$(P_{s,t}^\pm)^* (C \otimes M) P_{s,t}^\pm = C \otimes \left( \frac{1}{2} (M_{s,s} \pm M_{s,t} \pm M_{t,s} + M_{t,t}) \right).$$

Hence,

$$(P_{s,t}^+)^* (C \otimes M) P_{s,t}^+ - (P_{s,t}^-)^* (C \otimes M) P_{s,t}^- = C \otimes (M_{s,t} + M_{t,s}).$$

In particular, if  $M$  is symmetric, then the right hand side is  $2C \otimes M_{s,t}$ .

Let  $X \in \mathcal{D}_{L_A}(n)$  be given and let

$$Z = \sum_j A_j \otimes X_j.$$

By hypothesis,  $-X \in \mathcal{D}_{L_A}(n)$  too, so that both  $\pm Z \preceq I_n$ . Thus  $\pm(P_{s,t}^\pm)^* Z P_{s,t}^\pm \leq 1$ . Hence  $\pm(p_{s,t}^*)^\pm X p_{s,t}^\pm \in \mathcal{S}_{L_A}$  for each  $0 \leq s, t \leq n$ . Convexity of  $\mathcal{S}_{L_A}$  implies

$$\frac{1}{2}((p_{s,t}^+)^* X p_{s,t}^+ - (p_{s,t}^-)^* X p_{s,t}^-) = X_{s,t} := ((X_1)_{s,t}, \dots, (X_g)_{s,t}) \in \mathcal{S}_{L_A}.$$

By hypothesis,  $X_{s,t} \in \mathcal{S}_{L_B}$  and therefore,

$$T_{s,t} := \sum B_j (X_j)_{s,t} \preceq I_d, \quad 0 \leq s, t \leq n.$$

Apply Lemma 8.2 to the  $n \times n$  block matrix

$$T = \sum_j X_j \otimes B_j$$

to get

$$\|\sum B_j \otimes X_j\| \leq n$$

Likewise for  $-X$ , and therefore,

$$\sum B_j \otimes X_j \preceq nI_{dn}.$$

Hence  $\frac{1}{n}X \in \mathcal{D}_{L_B}$ . At this point we have  $\mathcal{D}_{L_A}(n) \subseteq n\mathcal{D}_{L_B}(n)$ .

Since  $B$  has size  $d$  and  $\mathcal{D}_{L_A}(d) \subseteq d\mathcal{D}_{L_B}(d)$ , it follows from Lemma 2.3 that  $\mathcal{D}_{L_A}(n) \subseteq d\mathcal{D}_{L_B}(n)$  for all  $n$ ; that is,  $\mathcal{D}_{L_A} \subseteq d\mathcal{D}_{L_B}$ .  $\blacksquare$

**Example 8.3.** This example shows that, in the case  $d = 2$ , the estimate  $r(A)(d)\mathcal{D}_{L_A} \subseteq \mathcal{D}_{L_B}(d)$  of Proposition 8.1 is sharp.

In this example  $\mathcal{S}_{L_A} = \mathcal{S}_{L_B}$  is the unit disc  $\mathbb{D} = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 \leq 1\}$ . We take  $L_A(X) \preceq 0$  to be the infinite set<sup>3</sup> of scalar inequalities

$$\sin(t)X_1 + \cos(t)X_2 \preceq I_n \quad \text{for all } t.$$

Next define  $L_B$  to be the pencil with coefficients

$$B_1 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \quad B_2 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

Of course  $d = \text{size}(L_B)$  is 2.

Now we show that  $\mathcal{D}_{L_A}(2) \not\subseteq (2 - \varepsilon)\mathcal{D}_{L_B}(2)$  for  $\varepsilon > 0$  by selecting  $X_j = B_j$ . Evidently,  $X \in \mathcal{D}_{L_A}(2)$  but, up to unitary equivalence,

$$L_B(X) = \begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & -1 & 1 & 0 \\ 0 & 1 & -1 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix}.$$

---

<sup>3</sup>For cp fans this actually is the the minimal operator system structure for  $\mathbb{D}$ .

Thus  $2I_4 - L_B(X) \succeq 0$ , but if  $\rho < 2$ , then  $\rho I_4 - L_B(X) \not\succeq 0$ .

To complete the example, we show that  $A$  can be viewed as a limit of tuples of matrices. Let  $\{t_j : j \in \mathbb{N}\}$  denote a countably dense subset of  $[0, 2\pi)$ . Given  $n \in \mathbb{N}$ , let  $T_n = \{t_1, \dots, t_n\}$  and let  $A_1^{(n)}$  denote the  $n \times n$  diagonal matrix with  $j$ -th diagonal entry  $\sin(t_j)$  and let  $A_2^{(n)}$  denote the  $n \times n$  diagonal matrix with  $j$ -th diagonal entry  $\cos(t_j)$ . In particular, if  $L_{A^{(n)}}(X) \succeq 0$  for all  $n$ , then  $L_A(X) \succeq 0$ . Thus, the smallest  $\rho$  such that  $L_{A^{(n)}}(X) \succeq 0$  implies  $L_B(\frac{1}{\rho}X) \succeq 0$  is 2.  $\square$

**8.2. The inclusion scale equals the commutability index.** The goal here is to prove Theorem 1.4 which we essentially restate as Theorem 8.4 and then prove.

Fix a tuple  $A \in \mathbb{S}_m^g$  and a positive integer  $d$ . Assume  $\mathcal{S}_{L_A} \subseteq \mathbb{R}^g$  is bounded. Let

$$\Omega_A(d) = \{r \geq 0 : \text{if } B \in \mathbb{S}_d^g \text{ and } \mathcal{S}_{L_A} \subseteq \mathcal{S}_{L_B}, \text{ then } r\mathcal{D}_{L_A} \subseteq \mathcal{D}_{L_B}\}.$$

Observe that  $\Omega \subseteq [0, 1]$ . Let  $\mathcal{F}_A$  denote the collection of tuples  $T = (T_1, \dots, T_g)$  of commuting self-adjoint operators on Hilbert space whose joint spectrum lies in  $\mathcal{S}_{L_A}$ . Let

$$\Gamma_A(d) = \{t \geq 0 : \text{if } X \in \mathcal{D}_{L_A}(d), \text{ then } tX \text{ dilates to a } T \in \mathcal{F}_A\}.$$

That  $\Gamma_A(d) \subseteq [0, 1]$  follows by noting that  $x$  is in the boundary of  $\mathcal{S}_{L_A}$  and if  $t > 1$ , then  $tx$  can not dilate to a  $T \in \mathcal{F}_A$ .

**Theorem 8.4.** *Fix  $g \in \mathbb{N}$ . Assuming  $\mathcal{S}_{L_A}$  is bounded, the sets  $\Gamma_A(d)$  and  $\Omega_A(d)$  contain non-zero positive numbers and are closed and equal. In particular, for each fixed  $d \in \mathbb{N}$ ,*

$$\sup \Omega_A(d) = \sup \Gamma_A(d).$$

The supremum of  $\Omega_A(d)$  is the optimal free spectrahedral inclusion constant for  $A$  and  $d$ . Namely, it is the largest number with the property that if  $B \in \mathbb{S}_d^g$  and  $\mathcal{S}_{L_A} \subseteq \mathcal{S}_{L_B}$ , then  $\Omega_A(d)\mathcal{D}_{L_A} \subseteq \mathcal{D}_{L_B}$ . On the other hand, the supremum of  $\Gamma_A(d)$  is the optimal scaling constant for  $A$  and  $d$  in the sense that if  $X \in \mathcal{D}_{L_A}(d)$  then  $\Gamma_A(d)X$  dilates to a tuple in  $\mathcal{F}_A$ .

**8.2.1. Matricial Hahn-Banach background.** The proof of this theorem given here uses the Effros-Winkler matricial Hahn-Banach Separation Theorem [EW97] for matrix convex sets. With  $g$  fixed let  $\mathbb{S}^g$  denote the sequence  $(\mathbb{S}_n^g)_n$ . A **matrix convex** subset  $\mathcal{C}$  of  $\mathbb{S}^g$  containing 0 is a sequence  $(\mathcal{C}(n))$  such that

- (a)  $\mathcal{C}(n) \subseteq \mathbb{S}_n^g$  for each  $n$ ;
- (b) 0 is in the interior of  $\mathcal{C}(1)$ ;
- (c)  $\mathcal{C}$  is **closed under direct sums**: If  $X \in \mathcal{C}(n)$  and  $Y \in \mathcal{C}(m)$ , then  $X \oplus Y = (X_1 \oplus Y_1, \dots, X_g \oplus Y_g) \in \mathcal{C}(n+m)$ , where

$$X_j \oplus Y_j = \begin{pmatrix} X_j & 0 \\ 0 & Y_j \end{pmatrix}.$$



(d)  $\mathcal{C}$  is **closed under simultaneous conjugation by contractions**: If  $X \in \mathcal{C}(n)$  and  $M$  is an  $n \times m$  contraction, then

$$M^*XM = (M^*X_1M, \dots, M^*X_gM) \in \mathcal{C}(m).$$

The matrix convex set  $\mathcal{C}$  is closed if each  $\mathcal{C}(n)$  is closed. A version of the matricial Hahn-Banach theorem immediately applicable here can be found in [HM12]. It says, in the language of this article, if  $\mathcal{C} \subseteq \mathbb{S}^g$  is closed and matrix convex and if  $X \in \mathbb{S}_d^g \setminus \mathcal{C}(d)$ , then there exists  $B \in \mathbb{S}_d^g$  such that  $\mathcal{C} \subseteq \mathcal{D}_{L_B}$ , but  $X \notin \mathcal{D}_{L_B}(d)$ . In particular,

$$\mathcal{C} = \bigcap \{ \mathcal{D}_{L_B} : B \in \mathbb{S}^g, \mathcal{C} \subseteq \mathcal{D}_{L_B} \}.$$

8.2.2. *Proof of Theorem 8.4.* For notational convenience, since  $A$  and  $d$  are fixed, let  $\Omega = \Omega_A(d)$ ,  $\Gamma = \Gamma_A(d)$  and  $\mathcal{F} = \mathcal{F}_A$ .

That  $\Omega$  is closed is easily seen. To see that  $\Omega$  contains a positive number, first note that the assumption that  $\mathcal{S}_{L_A}$  is bounded implies there exists a constant  $C > 0$  such that  $\mathcal{D}_{L_A} \subseteq C\mathfrak{C}^g$ . On the other hand, since  $\mathcal{D}_{L_A}$  is the set of tuples  $X \in \mathbb{S}^g$  such that  $I - \sum A_j \otimes X_j \succeq 0$ , there is a constant  $c > 0$  such that  $c\mathfrak{C}^g \subseteq \mathcal{D}_{L_A}$ . By Theorem 1.6 there is a constant  $s > 0$  such that if  $B \in \mathbb{S}_d^g$  and  $[-1, 1]^g \subseteq \mathcal{S}_{L_B}$ , then  $s\mathfrak{C}^g \subseteq \mathcal{D}_{L_B}$ . Hence, if instead  $\mathcal{S}_{L_A} \subseteq \mathcal{S}_{L_B}$ , then

$$c[-1, 1]^g \subseteq \mathcal{S}_{L_A} \subseteq \mathcal{S}_{L_B}.$$

It follows that  $\mathfrak{C}^g \subseteq \mathcal{D}_{\frac{B}{sc}}$ . Thus,

$$\mathcal{D}_{L_A} \subseteq C\mathfrak{C}^g \subseteq \frac{C}{sc}\mathcal{D}_{L_B}.$$

Hence  $\frac{sc}{C} \in \Omega$ .

To prove the sets  $\Omega$  and  $\Gamma$  are equal, first observe that Proposition 2.1 implies  $\Gamma \subseteq \Omega$ . To prove the converse, suppose  $r \in \Omega$ . Let  $\Sigma$  denote the smallest closed matrix convex set with the property that  $\Sigma(1) = \mathcal{D}_{L_A}(1) = \mathcal{S}_{L_A}$ . The equality,

$$\Sigma(d) = \bigcap \{ \mathcal{D}_{L_B}(d) : B \in \mathbb{S}_d^g \text{ and } \Sigma \subseteq \mathcal{D}_{L_B} \}$$

is a consequence of the Effros-Winkler matricial Hahn-Banach Separation Theorem [EW97]. To prove this assertion, first note the inclusion  $\Sigma(d)$  into the set on the right hand side is obvious. On the other hand, if  $X \notin \Sigma(d)$ , then by Effros-Winkler theorem produces a  $B \in \mathbb{S}_d^g$  such that  $\Sigma \subseteq \mathcal{D}_{L_B}$ , but  $X \notin \mathcal{D}_{L_B}(d)$  and the reverse inclusion follows. Now the definition of  $\Sigma$  implies  $\Sigma \subseteq \mathcal{D}_{L_B}$  if and only if  $\mathcal{S}_{L_A} = \Sigma(1) \subseteq \mathcal{S}_{L_B}$ . Hence,

$$\Sigma(d) = \bigcap \{ \mathcal{D}_{L_B}(d) : B \in \mathbb{S}_d^g \text{ and } \mathcal{S}_{L_A} \subseteq \mathcal{S}_{L_B} \}.$$

Thus, as  $\mathcal{S}_{L_A} \subseteq \mathcal{S}_{L_B}$  implies  $r\mathcal{D}_{L_A} \subseteq \mathcal{D}_{L_B}$ ,

$$\Sigma(d) \supseteq \bigcap \{ \mathcal{D}_{L_B}(d) : B \in \mathbb{S}_d^g \text{ and } r\mathcal{D}_{L_A} \subseteq \mathcal{D}_{L_B} \}$$

and therefore  $\Sigma(d) \supseteq r\mathcal{D}_{L_A}(d)$ .

It remains to show, if  $Z \in \Sigma$ , then  $Z$  dilates to some  $T \in \mathcal{F}$ . For positive integers  $n$ , let

$$\Lambda(n) = \{X \in \mathbb{S}_n^g : X \text{ dilates to some } T \in \mathcal{F}\}.$$

The sequence  $\Lambda = (\Lambda(n))_n$  is a matrix convex set with  $\Lambda(1) = \Sigma(1)$ . To prove that  $\Lambda(n)$  is closed, suppose  $(X^k)_k$  is a sequence from  $\Lambda(n)$  which converges to  $X \in \mathbb{S}_n^g$ . For each  $k$  there is a Hilbert space  $\mathcal{H}_k$ , a sequence of commuting self-adjoint contractions  $T^k = (T_1^k, \dots, T_g^k)$  on  $\mathcal{H}_k$  with joint spectrum in  $\mathcal{S}_{L_A}$  and an isometry  $V_k : \mathbb{R}^n \rightarrow \mathcal{H}_k$  such that

$$X^k = V_k^* T^k V_k.$$

Let  $T$  denote the tuple  $\oplus T^k$  acting on the Hilbert space  $\mathcal{H} = \oplus \mathcal{H}_k$ . The fact that each  $T^k$  has joint spectrum in the bounded set  $\mathcal{S}_{L_A}$  and that each  $T_j^k$  is self-adjoint, implies the sequence  $(T^k)_k$  is uniformly bounded. Hence  $T$  is a bounded operator. Let  $\mathcal{S}$  denote the operator system equal to the span of  $\{I, T_1, \dots, T_g\}$  (this set is self-adjoint since each  $T_k$  is self-adjoint) and let  $\phi : \mathcal{S} \rightarrow M_n$  denote the unital map determined by

$$\phi(T_j) = X_j.$$

It is straightforward to check that  $\phi$  is well defined. On the other hand, next it will be shown that  $\phi$  is completely positive, an argument which also shows that  $\phi$  is in fact well defined. If  $C = (C_0, \dots, C_g) \in \mathbb{S}_m^{g+1}$  and  $C_0 \otimes I + \sum C_j \otimes T_j \succeq 0$ , then  $C_0 \otimes I + \sum C_j \otimes T_j^k \succeq 0$  for each  $k$ . Thus,  $C_0 \otimes I + \sum C_j \otimes X_j^k \succeq 0$  for all  $k$  and finally  $C_0 \otimes I + \sum C_j \otimes X_j \succeq 0$ . Thus  $\phi$  is completely positive. Given a Hilbert space  $\mathcal{E}$ , let  $B(\mathcal{E})$  denote the C-star algebra of bounded operators on  $\mathcal{E}$ . By the standard application of Stinespring-Arveson ([Pau02, Corollary 7.7]) there exists a Hilbert space  $\mathcal{K}$ , a representation  $\pi : B(\mathcal{H}) \rightarrow B(\mathcal{K})$  and an isometry  $W : \mathbb{R}^n \rightarrow \mathcal{K}$  such that

$$X_j = \phi(T_j) = W^* \pi(T_j) W.$$

Since  $\pi$  is a representation, the tuple  $\pi(T) = (\pi(T_1), \dots, \pi(T_g))$  is a commuting tuple of self-adjoint contractions on the Hilbert space  $\mathcal{K}$  with joint spectrum in  $\mathcal{S}_{L_A}$ . Hence  $X \in \Lambda(n)$ .

Now  $\Lambda$  is a closed matrix convex set with  $\Lambda(1) \supseteq \Sigma(1)$ . Hence,  $\Sigma \subseteq \Lambda$  by the definition of  $\Sigma$ . In particular,  $r\mathcal{D}_{L_A}(d) \subseteq \Sigma(d) \subseteq \Lambda(d)$  and the proof is complete.  $\blacksquare$

**8.2.3. Matrix cube revisited.** Returning to the special case of the matrix cube, for  $g, d \in \mathbb{N}$  define

$$\rho_g(d) = \sup\{r \geq 0 : \text{if } B \in \mathbb{S}_d^g \text{ and } [-1, 1]^g \subseteq \mathcal{S}_{L_B}, \text{ then } r\mathfrak{C}^{(g)} \subseteq \mathcal{D}_{L_B}\}.$$

For  $d$  fixed, the sequence  $(\rho_g(d))_{g=1}^\infty$  is evidently decreasing and hence converges to some  $\rho(d)$ . Similarly, let  $\mathcal{F}_g$  denote the collection of tuples  $T = (T_1, \dots, T_g)$  of commuting self-adjoint contractions on Hilbert space and let

$$\tau_g(d) = \sup\{t \geq 0 : \text{if } X \in \mathfrak{C}^{(g)}(d), \text{ then } tX \text{ dilates to a } T \in \mathcal{F}_g\}.$$

The sequence  $(\tau_g(d))_{g=1}^\infty$  also decreases and hence converges to some  $\tau(d)$ . By Theorem 8.4,  $\tau_g(d) = \rho_g(d)$  for all  $g, d$ .

**Corollary 8.5.**  $\tau(d) = \lim \tau_g(d) = \lim \rho_g(d) = \frac{1}{\vartheta(d)}.$

**Remark 8.6.** To this point  $\tau(d) = \lim \rho_g(d)$  has been derived through operator theoretic means not involving  $\vartheta$  and [B-TN02]. Of course, in view of Theorems 1.6 and 1.7,  $\tau(d) = \frac{1}{\vartheta(d)}$ . On the other hand, it is not obviously possible to recover Theorem 1.1 from this corollary. See Remark 1.8.  $\square$

## 9. REFORMULATION OF THE OPTIMIZATION PROBLEM

The goal here is to bring pieces together in order to lay out our key classical optimization problem (1.1) in terms of regularized Beta functions (see Problem 9.1 and Proposition 9.2). The reformulated optimization problem is then solved in Section 12 after preliminary work in Sections 10 and 11.

Recall that  $\frac{1}{\vartheta(d)}$  for  $d \geq 2$  equals the minimum over all  $s, t \in \mathbb{N}$  and  $a, b \in \mathbb{R}_{>0}$  such that  $s + t = d = sa + tb$  of

$$2a \frac{s}{d} I_{\frac{a}{a+b}} \left( \frac{t}{2}, 1 + \frac{s}{2} \right) + 2b \frac{t}{d} I_{\frac{b}{a+b}} \left( \frac{s}{2}, 1 + \frac{t}{2} \right) - 1.$$

(Combine Lemma 6.6 with Proposition 4.2.) Note that the constraint  $d = sa + tb$  is just a matter of scaling of  $a$  and  $b$  with the same factor which won't affect the substitution

$$p = \frac{b}{a+b} \in (0, 1)$$

which we are now going to make. This substitution entails  $1 - p = \frac{a}{a+b}$ ,  $d = (a+b)(sp + t(1-p))$ ,

$$\frac{a}{d} = \frac{a}{(a+b)(sp + t(1-p))} = \frac{1-p}{sp + t(1-p)}$$

and

$$\frac{b}{d} = \frac{b}{(a+b)(sp + t(1-p))} = \frac{p}{sp + t(1-p)}.$$

By continuity, we can let  $p$  range over the compact interval  $[0, 1]$ . We therefore observe that  $\frac{1}{\vartheta(d)}$  equals the minimum over all  $s, t \in \mathbb{N}$  with  $s + t = d$  of the minimum value of the function  $f_{s,t}: [0, 1] \rightarrow \mathbb{R}$  given by

$$(9.1) \quad f_{s,t}(p) = \frac{2(1-p)sI_{1-p} \left( \frac{t}{2}, 1 + \frac{s}{2} \right) + 2ptI_p \left( \frac{s}{2}, 1 + \frac{t}{2} \right)}{(1-p)s + pt} - 1$$

for  $p \in [0, 1]$ . Using the standard identities  $I_p(x, y) = \frac{B_p(x, y)}{B(x, y)}$ ,  $\frac{\partial}{\partial p} B_p(x, y) = p^{x-1}(1-p)^{y-1}$ ,  $B \left( \frac{s}{2}, 1 + \frac{t}{2} \right) = \frac{t}{s+t} B \left( \frac{t}{2}, \frac{s}{2} \right)$  and  $B \left( \frac{t}{2}, 1 + \frac{s}{2} \right) = \frac{s}{s+t} B \left( \frac{s}{2}, \frac{t}{2} \right)$ , one can easily verify that the derivative  $f'_{s,t}$  of  $f_{s,t}$  takes the surprisingly simple form given by

$$f'_{s,t}(p) = \frac{2st}{((1-p)s + pt)^2} \left( I_p \left( \frac{s}{2}, 1 + \frac{t}{2} \right) - I_{1-p} \left( \frac{t}{2}, 1 + \frac{s}{2} \right) \right)$$

for  $p \in [0, 1]$  (two of the six terms cancel when one computes the derivative using the product and quotient rule). This shows that  $f_{s,t}$  is strictly decreasing on  $[0, \sigma_{s,t}]$  and strictly increasing on  $[\sigma_{s,t}, 1]$  where  $\sigma_{s,t} \in (0, 1)$  is defined by

$$(9.2) \quad I_{\sigma_{s,t}} \left( \frac{s}{2}, 1 + \frac{t}{2} \right) = I_{1-\sigma_{s,t}} \left( \frac{t}{2}, 1 + \frac{s}{2} \right).$$

We shall use (in Section 12) bounds on  $\sigma_{s,t}$  for  $s, t \in \mathbb{N}$ . Lower bounds are given in Corollary 12.5, while upper bounds are presented in Theorem 10.1, cf. (12.4).

**Problem 9.1.** Given a positive integer  $d$ , minimize

$$f_{s,t}(\sigma) = \frac{2(1-\sigma)sI_{1-\sigma}\left(\frac{t}{2}, 1 + \frac{s}{2}\right) + 2\sigma tI_{\sigma}\left(\frac{s}{2}, 1 + \frac{t}{2}\right)}{(1-\sigma)s + \sigma t} - 1$$

subject to the constraints

- (i)  $s, t \in \mathbb{N}$  and  $s + t = d$ ;
- (ii)  $s \geq \frac{d}{2}$ ;
- (iii)  $0 \leq \sigma \leq 1$ ; and
- (iv)  $I_{\sigma}\left(\frac{s}{2}, \frac{t}{2} + 1\right) = I_{1-\sigma}\left(\frac{t}{2}, \frac{s}{2} + 1\right)$ .

Since Problem 9.1 computes  $\vartheta(d)$ , Theorem 1.2 can be rephrased as follows.

**Proposition 9.2.** *When  $d$  is even the minimum in Problem 9.1 occurs when  $s = t = \frac{d}{2}$  and in this case  $\sigma = \frac{1}{2}$ . When  $d$  is odd, the minimum in Problem 9.1 occurs when  $s = \frac{d+1}{2}$  and  $t = \frac{d-1}{2}$ . In this case  $\sigma$ , and hence the optimum, is implicitly determined by condition (iv).*

The proof of Proposition 9.2 is organized as follows. The next section contains an improvement of Simmons' Theorem from probability. It is used to obtain the bound

$$\sigma_{s,t} \leq \frac{s}{s+t}$$

valid for  $s \geq \frac{d}{2}$ . In Section 11 we present the lower bound

$$\frac{s+2}{s+t+4} \leq \sigma_{s,t}$$

valid for  $s \geq \frac{d}{2}$ . Finally, the proof of Proposition 9.2 is completed in Section 12.

## 10. SIMMONS' THEOREM FOR HALF INTEGERS

This material has been motivated by the Perrin-Redside [PR07] proof of Simmons' inequality from discrete probability which has the following simple interpretation. Let  $s, d \in \mathbb{N}$  with  $s \geq \frac{d}{2}$ . Toss a coin whose probability for head is  $\frac{s}{d}$ ,  $d$  times. (So the expected number of head is  $s$ .) Simmons' inequality then states that the probability of getting  $< s$  heads is *smaller than* the probability of getting  $> s$  heads.

Theorem 10.1 below is a half-integer generalization of Simmons' Theorem.

**Theorem 10.1.** *For  $d \in \mathbb{N}$  and  $s, t \in \mathbb{N}$  with  $s + t = d$ , if  $\frac{d}{2} \leq s < d$ , then*

$$(10.1) \quad I_{\frac{s}{d}}\left(\frac{s}{2} + 1, \frac{t}{2}\right) \geq 1 - I_{\frac{s}{d}}\left(\frac{s}{2}, \frac{t}{2} + 1\right).$$

*Equivalently,*

$$(10.2) \quad \sigma_{s,t} \leq \frac{s}{d}$$

for  $\frac{d}{2} \leq s < d$  with  $s \in \mathbb{N}$ .

The proof of this consumes this whole section and we begin by setting notation. For  $s, t \in \mathbb{R}$  with  $s > -2$  and  $t > 0$  (in the sequel,  $s$  and  $t$  will mostly be integers with  $s \geq -1$  and  $t \geq 1$ ; we really need the case  $s = -1$ , for example after (10.15)) and  $d := s + t$ , let  $f_s$  denote the density function of the Beta distribution

$$\frac{d}{2} B\left(\frac{s}{2} + 1, \frac{t}{2}\right),$$

i.e.,

$$(10.3) \quad f_s(x) := \begin{cases} 0 & x \leq 0 \\ \frac{1}{B\left(\frac{s}{2} + 1, \frac{t}{2}\right)} \left(\frac{d}{2}\right)^{-1-\frac{s}{2}} x^{\frac{s}{2}} \left(1 - \frac{2}{d}x\right)^{\frac{t}{2}-1} & 0 < x < \frac{d}{2} \\ 0 & x \geq \frac{d}{2}. \end{cases}$$

Consider the function

$$\mathcal{F}(s, p) = I_p\left(\frac{s}{2} + 1, \frac{t}{2}\right) + I_p\left(\frac{s}{2}, \frac{t}{2} + 1\right) - 1.$$

Equation (10.1) is the statement that  $\mathcal{F}(s, \frac{s}{d}) \geq 0$ . Since, for  $s$  fixed,  $\mathcal{F}(s, p)$  strictly increases with  $p$  and  $\sigma_{s,t}$  is determined by  $\mathcal{F}(s, \sigma_{s,t}) = 0$ , the second part of the theorem is obviously equivalent to the first one.

Let

$$(10.4) \quad b_s := \int_0^{\frac{s}{2}} f_s = I_{\frac{s}{d}}\left(\frac{s}{2} + 1, \frac{t}{2}\right)$$

and

$$(10.5) \quad a_s := \int_{\frac{s}{2}}^{\infty} f_{s-2} = 1 - \int_0^{\frac{s}{2}} f_{s-2} = 1 - I_{\frac{s}{d}}\left(\frac{s}{2}, \frac{t}{2} + 1\right).$$

Equation (10.1) is equivalent to

$$(10.6) \quad c_s := b_s - a_s = \mathcal{F}\left(s, \frac{s}{d}\right) \geq 0$$

for  $d, s \in \mathbb{R}_{>0}$  with  $\frac{d}{2} \leq s < d$ .

**10.1. Two step monotonicity of  $c_s$ .** In this subsection, in Proposition 10.7, we show for  $s, d \in \mathbb{R}$  with  $\frac{d}{2} \leq s \leq d - 4$ , that  $c_{s+2} \geq c_s$ . Note that  $\frac{d}{2} \leq d - 4$  implies  $d \geq 8$ .

**Lemma 10.2.** *We have for  $s \in \mathbb{R}$  with  $0 < s < d - 2$ ,*

$$(10.7) \quad \begin{aligned} (xf_s)' &= \left(1 + \frac{s}{2}\right)(f_s - f_{s+2}) \\ ((d-2x)f_{s+2})' &= (d-s-2)(f_s - f_{s+2}). \end{aligned}$$

*Proof.* Straightforward. ■

For notational convenience we introduce, for  $s \in \mathbb{R}$  with  $-2 < s \leq d-3$ ,

$$\mathcal{I}_s := \int_{\frac{s}{2}}^{\frac{s}{2}+1} f_s.$$

**Lemma 10.3.** *For  $s \in \mathbb{R}$  with  $0 < s < d-2$ ,*

$$(10.8) \quad \begin{aligned} a_{s+2} - a_s &= f_s\left(\frac{s}{2}\right) - \mathcal{I}_s \\ b_{s+2} - b_s &= \mathcal{I}_s - f_s\left(\frac{s}{2} + 1\right). \end{aligned}$$

*Proof.* This is a consequence of the recursive formulas in Lemma 10.2:

$$\begin{aligned} a_{s+2} - a_s &= \int_{\frac{s}{2}+1}^{\infty} f_s - \int_{\frac{s}{2}}^{\infty} f_{s-2} \\ &= \int_{\frac{s}{2}}^{\infty} (f_s - f_{s-2}) - \int_{\frac{s}{2}}^{\frac{s}{2}+1} f_s \\ &\stackrel{(10.7)}{=} -\frac{d-2x}{d-s} f_s \Big|_{\frac{s}{2}}^{\infty} - \mathcal{I}_s \\ &= f_s\left(\frac{s}{2}\right) - \mathcal{I}_s. \end{aligned}$$

Similarly,

$$\begin{aligned} b_{s+2} - b_s &= \int_0^{\frac{s}{2}+1} f_{s+2} - \int_0^{\frac{s}{2}} f_s \\ &= \int_0^{\frac{s}{2}+1} (f_{s+2} - f_s) + \int_{\frac{s}{2}}^{\frac{s}{2}+1} f_s \\ &\stackrel{(10.7)}{=} -\frac{x f_s}{1 + \frac{s}{2}} \Big|_0^{\frac{s}{2}+1} + \mathcal{I}_s \\ &= \mathcal{I}_s - f_s\left(\frac{s}{2} + 1\right). \end{aligned} \quad \blacksquare$$

**Lemma 10.4.** *If  $s \in \mathbb{R}$  with  $-2 < s < d-2$ , then*

$$(10.9) \quad c_{s+2} - c_s = 2\mathcal{I}_s - f_s\left(\frac{s}{2}\right) - f_s\left(\frac{s}{2} + 1\right).$$

*Proof.* This is immediate from (10.6) and Lemma 10.3. \blacksquare

**Lemma 10.5.** *For  $0 < x < \frac{d}{2}$  and  $0 < s \leq d$ , the inequality  $f_s''(x) < 0$  holds if and only if*

$$(10.10) \quad \frac{4(d-4)(d-2)}{d^2} x^2 - \frac{4(d-4)s}{d} x + (s-2)s < 0.$$

*Proof.* Note that for  $0 < x < \frac{d}{2}$ ,

$$f_s''(x) = \frac{2^{\frac{s}{2}-1} d^{-s/2} x^{\frac{s}{2}-2} \left(1 - \frac{2x}{d}\right)^{\frac{d-s}{2}} \left(d^2(s-2)s - 4(d-4)dsx + 4(d-4)(d-2)x^2\right)}{(d-2x)^3 B\left(\frac{s+2}{2}, \frac{d-s}{2}\right)}.$$

Pulling a factor of  $d^2$  out of the last factor in the numerator yields (10.10). \blacksquare

**Lemma 10.6.** *If  $d, s \in \mathbb{R}$  and  $\frac{d}{2} \leq s \leq d-4$ , then  $f_s$  is concave on  $[\frac{s}{2}, \frac{s}{2} + 1]$ .*

*Proof.* Since  $s \geq \frac{d}{2}$  with  $s \leq d-4$ , then  $d \geq 8$  and thus  $s \geq 4$ . For very small  $x > 0$ , the left-hand side of (10.10) is positive and has a positive leading coefficient. So it suffices to verify (10.10) for  $x = \frac{s}{2}$  and  $x = \frac{s}{2} + 1$ . Let  $F_s(x)$  denote the left-hand side of (10.10). Then

$$\begin{aligned} F_s\left(\frac{s}{2}\right) &= \frac{2s}{d^2}(-d^2 + ds + 4s) \\ &\leq \frac{2s}{d^2}(-d^2 + d(d-4) + 4(d-4)) \\ &= -\frac{32s}{d^2} \\ &< 0. \end{aligned}$$

Similarly,

$$F_s\left(\frac{s}{2} + 1\right) = -\frac{2(d-s-2)}{d^2}(d(s-2) + 4(s+2)) < 0. \quad \blacksquare$$

**Proposition 10.7.** *For  $d, s \in \mathbb{R}$  with  $\frac{d}{2} \leq s \leq d-4$ , we have*

$$(10.11) \quad c_{s+2} > c_s.$$

*Furthermore,*

$$(10.12) \quad c_{\frac{d}{2}} = 0.$$

*Proof.* Since under the given constraints on  $s$ , the function  $f_s$  is concave on  $(\frac{s}{2}, \frac{s}{2} + 1)$ , its integral  $\mathcal{I}_s$  over this interval is bigger than

$$\frac{1}{2}\left(f_s\left(\frac{s}{2}\right) + f_s\left(\frac{s}{2} + 1\right)\right).$$

The Equation (10.11) now follows from Lemma 10.4.

Using  $I_x(a, b) = 1 - I_{1-x}(b, a)$  we get that

$$a_{\frac{d}{2}} = 1 - I_{\frac{1}{2}}\left(\frac{d}{4}, \frac{d}{4} + 1\right) = I_{\frac{1}{2}}\left(\frac{d}{4} + 1, \frac{d}{4}\right) = b_{\frac{d}{2}},$$

whence  $c_{\frac{d}{2}} = 0$ . \blacksquare

For  $d \geq 8$  and even, an implication of two step monotonicity in Proposition 10.7 together with (10.12) is that either

$$\min\{c_s : \frac{d}{2} \leq s < d\} = c_{d-1}$$

or

$$(10.13) \quad \min\{c_s : \frac{d}{2} \leq s < d, s \text{ even}\} = c_{\frac{d}{2}} = 0 \quad \text{and} \quad \min\{c_s : \frac{d}{2} \leq s < d, s \text{ odd}\} = c_{\frac{d}{2}+1}.$$

Likewise for  $d \geq 8$  and odd either

$$\min\{c_s : \frac{d}{2} \leq s < d\} = c_{d-1}$$

or

$$(10.14) \quad \min_{s \text{ even}} c_s = c_{\frac{d}{2}+1} \quad \text{and} \quad \min_{s \text{ odd}} c_s = c_{\frac{d}{2}+\frac{3}{2}}.$$

Thus proving Theorem 10.1 for an even integer  $d \geq 8$  reduces to establishing that  $c_{d-1} > 0$  and  $c_{\frac{d}{2}+1}$  are nonnegative and proving the result for  $d \geq 8$  odd reduces to showing in addition that  $c_{\frac{d}{2}+1}$  and  $c_{\frac{d}{2}+\frac{3}{2}}$  are both nonnegative, facts established in Sections 10.3 and 10.4 respectively. Finally, that  $c_s \geq 0$  for all  $d, s$  with  $\frac{d}{2} \leq s \leq d < 8$  was checked symbolically by Mathematica.

**10.2. The upper boundary case.** This subsection is devoted to proving the following lemma which is essential for proving both the even and odd cases of Proposition 10.7.

**Lemma 10.8.**  $c_{d-1} > 0$  for  $d \in \mathbb{N}$  with  $d \geq 2$ . In addition if  $d \in \mathbb{R}$  with  $d \geq 6$  then  $c_{d-1} > 0$ .

*Proof.* It is easy to verify the given inequality by hand for  $d = 2, 3, 4, 5$ . Without loss of generality we assume  $d \in \mathbb{R}$  with  $d \geq 6$  in what follows.

Recall that  $c_s = b_s - a_s$ . Observe that

$$b_{d-1} = \int_0^{\frac{d-1}{2}} f_{d-1} = 1 - \int_{\frac{d-1}{2}}^{\frac{d}{2}} f_{d-1},$$

and recall that,

$$a_{d-1} = \int_{\frac{d-1}{2}}^{\frac{d}{2}} f_{d-3}.$$

Hence  $c_{d-1} \geq 0$  iff

$$(10.15) \quad \int_{\frac{d-1}{2}}^{\frac{d}{2}} (f_{d-1} + f_{d-3}) \leq 1.$$

We next use Lemma 10.2 with  $s = d - 3 \geq -1 > -2$  to express

$$f_{d-1} = f_{d-3} - ((d-2x)f_{d-1})',$$

whence the left-hand side of (10.15) transforms into

$$(10.16) \quad \int_{\frac{d-1}{2}}^{\frac{d}{2}} (f_{d-1} + f_{d-3}) = 2 \int_{\frac{d-1}{2}}^{\frac{d}{2}} f_{d-3} + f_{d-1} \left( \frac{d-1}{2} \right).$$

**Sublemma 10.9.**  $f_{d-3}$  is concave on  $(\frac{d-1}{2}, \frac{d}{2})$ .

*Proof.* First note that

$$f_{d-3}''(x) = \frac{2^{\frac{d-5}{2}} d^{\frac{1}{2}} (-d-1) x^{\frac{d-7}{2}} ((d-5)(d-3)d^2 + 4(d-4)(d-2)x^2 - 4(d-4)(d-3)dx)}{(d-2x)\sqrt{1 - \frac{2x}{d}B\left(\frac{d-1}{2}, \frac{3}{2}\right)}}$$



by a straightforward computation. Thus the sign of  $f''_{d-3}(x)$  is determined by that of

$$\begin{aligned} (d-5)(d-3)d^2 - 4(d-4)(d-3)d x + 4(d-4)(d-2)x^2 \\ = 4(d-4)(d-2) \left( x - \frac{(d-3)d}{2(d-2)} \right)^2 - \frac{2(d-3)d^2}{d-2} \end{aligned}$$

Since  $d > 5$  and  $\frac{(d-3)d}{2(d-2)} \leq \frac{d-1}{2} \leq x \leq \frac{d}{2}$ , this expression can be bound as follows:

$$\begin{aligned} (d-5)(d-3)d^2 - 4(d-4)(d-3)d x + 4(d-4)(d-2)x^2 \\ \leq 4(d-4)(d-2) \left( \frac{d}{2} - \frac{(d-3)d}{2(d-2)} \right)^2 - \frac{2(d-3)d^2}{d-2} = -d^2 < 0. \quad \blacksquare \end{aligned}$$

By Sublemma 10.9,

$$2 \int_{\frac{d-1}{2}}^{\frac{d}{2}} f_{d-3} \leq f_{d-3} \left( \frac{d}{2} - \frac{1}{4} \right).$$

It thus suffices to establish

$$(10.17) \quad f_{d-3} \left( \frac{d}{2} - \frac{1}{4} \right) + f_{d-1} \left( \frac{d}{2} - \frac{1}{2} \right) \leq 1.$$

The left-hand side of (10.17) expands into

$$\Psi(d) := \frac{2d^{1-\frac{d}{2}} \left( (d-1)^{\frac{d-3}{2}} + 2^{1-\frac{d}{2}}(2d-1)^{\frac{d-3}{2}} \right) \Gamma\left(\frac{d}{2}\right)}{\sqrt{\pi} \Gamma\left(\frac{d-1}{2}\right)}.$$

We use [KV71, (1.7)]:

$$\frac{\Gamma\left(\frac{d}{2}\right)}{\Gamma\left(\frac{d-1}{2}\right)} \leq \frac{d^{\frac{d}{2}-\frac{1}{2}}}{\sqrt{2e} (d-1)^{\frac{d}{2}-1}}.$$

Using this,  $\Psi(d) \leq 1$  will follow once we establish

$$\sqrt{\frac{\pi e}{2d}} (d-1)^{\frac{d}{2}-1} \geq (d-1)^{\frac{d}{2}-\frac{3}{2}} + \frac{1}{\sqrt{2}} \left( d - \frac{1}{2} \right)^{\frac{d}{2}-\frac{3}{2}},$$

i.e.,

$$(10.18) \quad \sqrt{\frac{\pi e}{2}} \geq \sqrt{d} \left( (d-1)^{-\frac{1}{2}} + \frac{1}{\sqrt{2}} \left( d - \frac{1}{2} \right)^{-\frac{1}{2}} \left( 1 + \frac{1}{2(d-1)} \right)^{\frac{d}{2}-1} \right).$$

**Sublemma 10.10.** The sequence

$$(10.19) \quad \left( 1 + \frac{1}{2(d-1)} \right)^{\frac{d}{2}-1}$$

is increasing with limit

$$\sqrt[4]{e}.$$

*Proof.* Letting  $\delta = 2(d-1)$ , (10.19) can be rephrased as

$$\left(1 + \frac{1}{\delta}\right)^{\frac{\delta}{4}} \left(1 + \frac{1}{\delta}\right)^{-\frac{1}{2}}.$$

The first factor is often used to define  $e$  and is well-known to form an increasing sequence. It is clear that the sequence  $(1 + \frac{1}{\delta})^{-\frac{1}{2}}$  is increasing.  $\blacksquare$

Now the right-hand side of (10.18) can be bound above by

$$(10.20) \quad \sqrt{d} \left( (d-1)^{-\frac{1}{2}} + \frac{1}{\sqrt{2}} \left( d - \frac{1}{2} \right)^{-\frac{1}{2}} \left( 1 + \frac{1}{2(d-1)} \right)^{\frac{d}{2}-1} \right) \\ \leq \sqrt{d} \left( (d-1)^{-\frac{1}{2}} + \frac{\sqrt[4]{e}}{\sqrt{2}} \left( d - \frac{1}{2} \right)^{-\frac{1}{2}} \right) =: \Phi(d).$$

Both of the sequences

$$\sqrt{d}(d-1)^{-\frac{1}{2}} \quad \text{and} \quad \sqrt{d} \left( d - \frac{1}{2} \right)^{-\frac{1}{2}}$$

are decreasing, and the limit as  $d \rightarrow \infty$  of the right-hand side of (10.20) is

$$1 + \frac{\sqrt[4]{e}}{\sqrt{2}} \approx 1.90794.$$

Since

$$\Phi(6) = \sqrt{\frac{6}{5}} + \sqrt{\frac{6}{11}} \sqrt[4]{e} \leq \sqrt{\frac{\pi e}{2}},$$

all this establishes (10.17) for  $d \geq 6$  as was required.  $\blacksquare$

**10.3. The lower boundary cases for  $d$  even.** Here we prove  $c_{\frac{d}{2}+1} > 0$  for  $d \in \mathbb{R}$  with  $d \geq 2$  (the other lower boundary case,  $c_{\frac{d}{2}} \geq 0$  was already proved), thus establishing (10.13) and proving Theorem 10.1 for  $d$  even.

**Lemma 10.11.**  $c_{\frac{d}{2}+1} > 0$  for  $d \in \mathbb{R}$  such that  $d \geq 2$ .

*Proof.* We want to prove that

$$(10.21) \quad I_{\frac{1}{2}+\frac{1}{d}} \left( \frac{d}{4} + \frac{3}{2}, \frac{d}{4} - \frac{1}{2} \right) + I_{\frac{1}{2}+\frac{1}{d}} \left( \frac{d}{4} + \frac{1}{2}, \frac{d}{4} + \frac{1}{2} \right) \geq 1.$$

Using

$$(10.22) \quad I_x(a+1, b-1) = I_x(a, b) - \frac{x^a(1-x)^{b-1}}{a B(a, b)}$$

we rewrite the first summand in (10.21) as

$$(10.23) \quad I_{\frac{1}{2}+\frac{1}{d}} \left( \frac{d}{4} + \frac{3}{2}, \frac{d}{4} - \frac{1}{2} \right) = I_{\frac{1}{2}+\frac{1}{d}} \left( \frac{d}{4} + \frac{1}{2}, \frac{d}{4} + \frac{1}{2} \right) - \frac{\left(\frac{1}{2} + \frac{1}{d}\right)^{\frac{1}{2}+\frac{d}{4}} \left(\frac{1}{2} - \frac{1}{d}\right)^{-\frac{1}{2}+\frac{d}{4}}}{\left(\frac{d}{4} + \frac{1}{2}\right) B\left(\frac{d}{4} + \frac{1}{2}, \frac{d}{4} + \frac{1}{2}\right)}.$$

Upon multiplying (10.21) with  $B\left(\frac{d}{4} + \frac{1}{2}, \frac{d}{4} + \frac{1}{2}\right)$  and using (10.23), (10.21) is equivalent to

$$(10.24) \quad 2B_{\frac{1}{2}+\frac{1}{d}}\left(\frac{d}{4} + \frac{1}{2}, \frac{d}{4} + \frac{1}{2}\right) \geq B\left(\frac{d}{4} + \frac{1}{2}, \frac{d}{4} + \frac{1}{2}\right) + \frac{\left(\frac{1}{2} + \frac{1}{d}\right)^{\frac{1}{2}+\frac{d}{4}} \left(\frac{1}{2} - \frac{1}{d}\right)^{-\frac{1}{2}+\frac{d}{4}}}{\frac{d}{4} + \frac{1}{2}}.$$

The left-hand side of this inequality can be rewritten as

$$\begin{aligned} 2B_{\frac{1}{2}+\frac{1}{d}}\left(\frac{d}{4} + \frac{1}{2}, \frac{d}{4} + \frac{1}{2}\right) &= 2 \int_0^{\frac{1}{2}+\frac{1}{d}} x^{\frac{d}{4}-\frac{1}{2}} (1-x)^{\frac{d}{4}-\frac{1}{2}} dx \\ &= 2B_{\frac{1}{2}}\left(\frac{d}{4} + \frac{1}{2}, \frac{d}{4} + \frac{1}{2}\right) + 2 \int_{\frac{1}{2}}^{\frac{1}{2}+\frac{1}{d}} x^{\frac{d}{4}-\frac{1}{2}} (1-x)^{\frac{d}{4}-\frac{1}{2}} dx \\ &= B\left(\frac{d}{4} + \frac{1}{2}, \frac{d}{4} + \frac{1}{2}\right) + 2 \int_{\frac{1}{2}}^{\frac{1}{2}+\frac{1}{d}} x^{\frac{d}{4}-\frac{1}{2}} (1-x)^{\frac{d}{4}-\frac{1}{2}} dx. \end{aligned}$$

Now (10.24) is equivalent to

$$(10.25) \quad 2 \int_{\frac{1}{2}}^{\frac{1}{2}+\frac{1}{d}} x^{\frac{d}{4}-\frac{1}{2}} (1-x)^{\frac{d}{4}-\frac{1}{2}} dx \geq \frac{\left(\frac{1}{2} + \frac{1}{d}\right)^{\frac{1}{2}+\frac{d}{4}} \left(\frac{1}{2} - \frac{1}{d}\right)^{-\frac{1}{2}+\frac{d}{4}}}{\frac{d}{4} + \frac{1}{2}}.$$

The derivative of the integrand on the left hand side equals

$$-\frac{1}{4}(d-2)(2x-1)((1-x)x)^{\frac{d-6}{4}}$$

and is thus nonpositive on  $[\frac{1}{2}, 1] \supseteq [\frac{1}{2}, \frac{1}{2} + \frac{1}{d}]$ . Hence

$$\begin{aligned} 2 \int_{\frac{1}{2}}^{\frac{1}{2}+\frac{1}{d}} x^{\frac{d}{4}-\frac{1}{2}} (1-x)^{\frac{d}{4}-\frac{1}{2}} dx &\geq \frac{2}{d} \left(\frac{1}{2} + \frac{1}{d}\right)^{\frac{d}{4}-\frac{1}{2}} \left(\frac{1}{2} - \frac{1}{d}\right)^{\frac{d}{4}-\frac{1}{2}} \\ &= \frac{\left(\frac{1}{2} + \frac{1}{d}\right)^{\frac{1}{2}+\frac{d}{4}} \left(\frac{1}{2} - \frac{1}{d}\right)^{-\frac{1}{2}+\frac{d}{4}}}{\frac{d}{4} + \frac{1}{2}}, \end{aligned}$$

as desired. ■

This concludes the proof of Theorem 10.1 for even  $d$ .

**10.4. The lower boundary cases for  $d$  odd.** In this subsection we establish two key inequalities:

$$c_{\frac{d}{2}+\frac{1}{2}} > 0 \quad \text{and} \quad c_{\frac{d}{2}+\frac{3}{2}} > 0.$$

We show that the first holds for  $d \in \mathbb{R}$  with  $d \geq 3$  and the second for  $d \in \mathbb{N}$  with  $d \geq 3$  or for  $d \in \mathbb{R}$  with  $d \geq 16$ . Combined with Lemma 10.8 these inequalities show that the minimum of  $c_s$  over  $\frac{d}{2} \leq s \leq d-1$  is strictly positive and as a consequence proves Theorem 10.1 in the case of  $d$  odd.

**Lemma 10.12.**  $c_{\frac{d}{2}+\frac{1}{2}} > 0$  for  $d \in \mathbb{R}$  and  $d \geq 3$ .

*Proof.* We claim that

$$(10.26) \quad I_{\frac{1}{2}+\frac{1}{2d}} \left( \frac{d}{4} + \frac{5}{4}, \frac{d}{4} - \frac{1}{4} \right) + I_{\frac{1}{2}+\frac{1}{2d}} \left( \frac{d}{4} + \frac{1}{4}, \frac{d}{4} + \frac{3}{4} \right) \geq 1.$$

Using (10.22) we rewrite the first summand in (10.26) as

$$(10.27) \quad I_{\frac{1}{2}+\frac{1}{2d}} \left( \frac{d}{4} + \frac{5}{4}, \frac{d}{4} - \frac{1}{4} \right) = I_{\frac{1}{2}+\frac{1}{2d}} \left( \frac{d}{4} + \frac{1}{4}, \frac{d}{4} + \frac{3}{4} \right) - \frac{\left(\frac{1}{2} + \frac{1}{2d}\right)^{\frac{1}{4}+\frac{d}{4}} \left(\frac{1}{2} - \frac{1}{2d}\right)^{-\frac{1}{4}+\frac{d}{4}}}{\left(\frac{d}{4} + \frac{1}{4}\right) B\left(\frac{d}{4} + \frac{1}{4}, \frac{d}{4} + \frac{3}{4}\right)}.$$

Upon multiplying (10.26) with  $B\left(\frac{d}{4} + \frac{1}{4}, \frac{d}{4} + \frac{3}{4}\right)$  and using (10.27), (10.26) rewrites to

$$(10.28) \quad 2B_{\frac{1}{2}+\frac{1}{2d}} \left( \frac{d}{4} + \frac{1}{4}, \frac{d}{4} + \frac{3}{4} \right) \geq B\left(\frac{d}{4} + \frac{1}{4}, \frac{d}{4} + \frac{3}{4}\right) + \frac{4\left(1 + \frac{1}{d}\right)^{\frac{1}{4}+\frac{d}{4}} \left(1 - \frac{1}{d}\right)^{-\frac{1}{4}+\frac{d}{4}}}{2^{\frac{d}{2}}(d+1)}$$

The left-hand side of this inequality can be expanded as

$$\begin{aligned} 2B_{\frac{1}{2}+\frac{1}{2d}} \left( \frac{d}{4} + \frac{1}{4}, \frac{d}{4} + \frac{3}{4} \right) &= 2 \int_0^{\frac{1}{2}+\frac{1}{2d}} x^{\frac{d}{4}-\frac{3}{4}} (1-x)^{\frac{d}{4}-\frac{1}{4}} dx \\ &= 2B_{\frac{1}{2}} \left( \frac{d}{4} + \frac{1}{4}, \frac{d}{4} + \frac{3}{4} \right) + 2 \int_{\frac{1}{2}}^{\frac{1}{2}+\frac{1}{2d}} x^{\frac{d}{4}-\frac{3}{4}} (1-x)^{\frac{d}{4}-\frac{1}{4}} dx. \end{aligned}$$

Now (10.28) is equivalent to

$$\begin{aligned} (10.29) \quad 2B_{\frac{1}{2}} \left( \frac{d}{4} + \frac{1}{4}, \frac{d}{4} + \frac{3}{4} \right) &+ 2 \int_{\frac{1}{2}}^{\frac{1}{2}+\frac{1}{2d}} x^{\frac{d}{4}-\frac{3}{4}} (1-x)^{\frac{d}{4}-\frac{1}{4}} dx \\ &\geq B\left(\frac{d}{4} + \frac{1}{4}, \frac{d}{4} + \frac{3}{4}\right) + \frac{4\left(1 + \frac{1}{d}\right)^{\frac{1}{4}+\frac{d}{4}} \left(1 - \frac{1}{d}\right)^{-\frac{1}{4}+\frac{d}{4}}}{2^{\frac{d}{2}}(d+1)}. \end{aligned}$$

A brief calculation shows

$$(10.30) \quad 2B_{\frac{1}{2}} \left( \frac{d}{4} + \frac{1}{4}, \frac{d}{4} + \frac{3}{4} \right) - B\left(\frac{d}{4} + \frac{1}{4}, \frac{d}{4} + \frac{3}{4}\right) = \int_0^{1/2} x^{\frac{d}{4}-\frac{3}{4}} (1-x)^{\frac{d}{4}-\frac{3}{4}} (\sqrt{1-x} - \sqrt{x}) dx.$$

We next set out to provide a lower bound on (10.30). First, rewrite

$$(10.31) \quad \sqrt{1-x} - \sqrt{x} = \frac{1-2x}{\sqrt{1-x} + \sqrt{x}}$$

and observe that  $\sqrt{1-x} + \sqrt{x}$  is increasing from 1 to  $\sqrt{2}$  on  $[0, \frac{1}{2}]$ . Hence (10.31) can be bound as

$$(10.32) \quad \frac{1}{\sqrt{2}}(1-2x) \leq \sqrt{1-x} - \sqrt{x} \leq 1-2x.$$

This gives

$$\begin{aligned}
 (10.30) &\geq \frac{1}{\sqrt{2}} \int_0^{1/2} x^{\frac{d}{4}-\frac{3}{4}} (1-x)^{\frac{d}{4}-\frac{3}{4}} (1-2x) dx \\
 (10.33) &= \frac{1}{\sqrt{2}} \int_0^{1/2} x^{\frac{d}{4}-\frac{3}{4}} (1-x)^{\frac{d}{4}-\frac{3}{4}} dx - \sqrt{2} \int_0^{1/2} x^{\frac{d}{4}+\frac{1}{4}} (1-x)^{\frac{d}{4}-\frac{3}{4}} dx \\
 &= \frac{1}{\sqrt{2}} B_{\frac{1}{2}} \left( \frac{d}{4} + \frac{1}{4}, \frac{d}{4} + \frac{1}{4} \right) - \sqrt{2} B_{\frac{1}{2}} \left( \frac{d}{4} + \frac{5}{4}, \frac{d}{4} + \frac{1}{4} \right).
 \end{aligned}$$

Using the well-known formula

$$B(z, a+1, b) = \frac{aB(z, a, b) - z^a(1-z)^b}{a+b}$$

on  $B_{\frac{1}{2}} \left( \frac{d}{4} + \frac{5}{4}, \frac{d}{4} + \frac{1}{4} \right)$ , the final expression in (10.33) simplifies into

$$(10.34) \quad \frac{2}{2^{\frac{d}{2}}(d+1)}.$$

**Sublemma 10.13.** For  $3 \leq d \in \mathbb{R}$ , the integrand  $\eta(x)$  in the second summand in (10.29) is decreasing and concave on  $(\frac{1}{2}, \frac{1}{2} + \frac{1}{2d})$ .

*Proof.* This is routine. The derivative of  $\eta$  is

$$\eta'(x) = \frac{1}{4} (1-x)^{\frac{d-5}{4}} x^{\frac{d-7}{4}} (-2dx + 4x + d - 3)$$

So its sign on  $(\frac{1}{2}, \frac{1}{2} + \frac{1}{2d})$  is governed by that of  $-2dx + 4x + d - 3$ . However,

$$-2dx + 4x + d - 3 \leq 2d \cdot \frac{1}{2} + 4 \left( \frac{1}{2} + \frac{1}{2d} \right) + d - 3 = -1 + \frac{2}{d}$$

is negative for  $d \geq 3$ . Thus  $\eta$  is decreasing.

Further,

$$\eta''(x) = \frac{1}{16} (1-x)^{\frac{d-9}{4}} x^{\frac{d-11}{4}} (4(d-4)(d-2)x^2 - 4(d-4)(d-3)x + d^2 - 10d + 21).$$

For  $\frac{1}{2} \leq x \leq \frac{1}{2} + \frac{1}{2d}$  we have,

$$\begin{aligned}
 &4(d-4)(d-2)x^2 - 4(d-4)(d-3)x + d^2 - 10d + 21 \\
 &\leq 4(d-4)(d-2) \left( \frac{1}{2} + \frac{1}{2d} \right)^2 - 4(d-4)(d-3) \cdot \frac{1}{2} + d^2 - 10d + 21 \\
 &= \frac{8}{d^2} + \frac{10}{d} - 6
 \end{aligned}$$

and the last expression is negative for  $d \geq 3$ , whence  $\eta''(x) < 0$ . ■

By Sublemma 10.13, the integral in (10.29) can be bound below by

$$\frac{1}{2} \frac{1}{2d} \left( \left( \frac{1}{2} + \frac{1}{2d} \right)^{\frac{d}{4}-\frac{3}{4}} \left( \frac{1}{2} - \frac{1}{2d} \right)^{\frac{d}{4}-\frac{1}{4}} + 2^{1-\frac{d}{2}} \right).$$

Moving this to the right-hand side of (10.29) means we have to show that (10.30) is at least

$$(10.35) \quad \frac{1}{2^{\frac{d}{2}} d} \left( 3 \left( 1 + \frac{1}{d} \right)^{\frac{d-3}{4}} \left( 1 - \frac{1}{d} \right)^{\frac{d-1}{4}} - 1 \right).$$

It suffices to replace (10.30) with its lower bound (10.34). That is, we shall prove

$$(10.36) \quad \frac{3d+1}{d+1} \geq 3 \left( 1 + \frac{1}{d} \right)^{\frac{d-3}{4}} \left( 1 - \frac{1}{d} \right)^{\frac{d-1}{4}}.$$

Rearranging the right-hand side of (10.36) we get

$$\begin{aligned} 3 \left( 1 + \frac{1}{d} \right)^{\frac{d-3}{4}} \left( 1 - \frac{1}{d} \right)^{\frac{d-1}{4}} &= 3 \left( 1 + \frac{1}{d} \right)^{\frac{d+1}{4}} \left( 1 - \frac{1}{d} \right)^{\frac{d-1}{4}} \left( 1 + \frac{1}{d} \right)^{-1} \\ &= 3 \left( 1 + \frac{1}{d} \right)^{\frac{d+1}{4}} \left( 1 - \frac{1}{d} \right)^{\frac{d-1}{4}} \frac{d}{d+1}. \end{aligned}$$

In particular, (10.36) is equivalent to

$$(10.37) \quad 3d+1 \geq 3d \left( 1 + \frac{1}{d} \right)^{\frac{d+1}{4}} \left( 1 - \frac{1}{d} \right)^{\frac{d-1}{4}}.$$

As before, the sequence  $\left( 1 + \frac{1}{d} \right)^{\frac{d+1}{4}}$  is increasing with limit  $e^{\frac{1}{4}}$ , so

$$\begin{aligned} (10.38) \quad 3d \left( 1 + \frac{1}{d} \right)^{\frac{d+1}{4}} \left( 1 - \frac{1}{d} \right)^{\frac{d-1}{4}} &\leq 3de^{\frac{1}{4}} \left( 1 - \frac{1}{d} \right)^{\frac{d-1}{4}} \\ &= 3de^{\frac{1}{4}} \left( 1 - \frac{1}{d} \right)^{\frac{d}{4}} \left( 1 - \frac{1}{d} \right)^{-\frac{1}{4}} \end{aligned}$$

The sequence  $\left( 1 - \frac{1}{d} \right)^{\frac{d}{4}}$  is increasing with limit  $e^{-\frac{1}{4}}$ , so the right-hand side of (10.38) is further at most

$$3d \left( 1 - \frac{1}{d} \right)^{-\frac{1}{4}}.$$

Now (10.37) is implied by

$$1 + \frac{1}{3d} \geq \left( 1 - \frac{1}{d} \right)^{-\frac{1}{4}},$$

an inequality easy to establish using calculus. ■

**Lemma 10.14.**  $c_{\frac{d}{2}+\frac{3}{2}} > 0$  for  $3 \leq d \in \mathbb{N}$ . In addition, if  $d \in \mathbb{R}$  and  $d \geq 16$ , then  $c_{\frac{d}{2}+\frac{3}{2}} > 0$ .

*Proof.* We claim that

$$(10.39) \quad I_{\frac{1}{2}+\frac{3}{2d}} \left( \frac{d}{4} + \frac{7}{4}, \frac{d}{4} - \frac{3}{4} \right) + I_{\frac{1}{2}+\frac{3}{2d}} \left( \frac{d}{4} + \frac{3}{4}, \frac{d}{4} + \frac{1}{4} \right) \geq 1.$$

Using (10.22) we rewrite the first summand in (10.39) as

$$(10.40) \quad I_{\frac{1}{2}+\frac{3}{2d}}\left(\frac{d}{4}+\frac{7}{4}, \frac{d}{4}-\frac{3}{4}\right) = I_{\frac{1}{2}+\frac{3}{2d}}\left(\frac{d}{4}+\frac{3}{4}, \frac{d}{4}+\frac{1}{4}\right) - \frac{\left(\frac{1}{2}+\frac{3}{2d}\right)^{\frac{3}{4}+\frac{d}{4}}\left(\frac{1}{2}-\frac{3}{2d}\right)^{-\frac{3}{4}+\frac{d}{4}}}{\left(\frac{d}{4}+\frac{3}{4}\right)B\left(\frac{d}{4}+\frac{3}{4}, \frac{d}{4}+\frac{1}{4}\right)}.$$

Upon multiplying (10.39) with  $B\left(\frac{d}{4}+\frac{3}{4}, \frac{d}{4}+\frac{1}{4}\right)$  and using (10.40), (10.39) rewrites to

$$(10.41) \quad 2B_{\frac{1}{2}+\frac{3}{2d}}\left(\frac{d}{4}+\frac{3}{4}, \frac{d}{4}+\frac{1}{4}\right) \geq B\left(\frac{d}{4}+\frac{3}{4}, \frac{d}{4}+\frac{1}{4}\right) + \frac{4\left(1+\frac{3}{d}\right)^{\frac{3}{4}+\frac{d}{4}}\left(1-\frac{3}{d}\right)^{-\frac{3}{4}+\frac{d}{4}}}{2^{\frac{d}{2}}(d+3)}$$

Further, using

$$(10.42) \quad B_z(a+1, b) = \frac{aB_z(a, b) - z^a(1-z)^b}{a+b}$$

on the two betas in (10.41), we get

$$\begin{aligned} \frac{d-1}{d}B_{\frac{1}{2}+\frac{3}{2d}}\left(\frac{d}{4}-\frac{1}{4}, \frac{d}{4}+\frac{1}{4}\right) - \frac{2\left(1+\frac{3}{d}\right)^{-\frac{1}{4}+\frac{d}{4}}\left(1-\frac{3}{d}\right)^{+\frac{1}{4}+\frac{d}{4}}}{2^{\frac{d}{2}}d} \\ \geq \frac{d-1}{2d}B\left(\frac{d}{4}-\frac{1}{4}, \frac{d}{4}+\frac{1}{4}\right) + \frac{4\left(1+\frac{3}{d}\right)^{\frac{3}{4}+\frac{d}{4}}\left(1-\frac{3}{d}\right)^{-\frac{3}{4}+\frac{d}{4}}}{2^{\frac{d}{2}}(d+3)}, \end{aligned}$$

or equivalently,

$$(10.43) \quad 2B_{\frac{1}{2}+\frac{3}{2d}}\left(\frac{d}{4}-\frac{1}{4}, \frac{d}{4}+\frac{1}{4}\right) - B\left(\frac{d}{4}-\frac{1}{4}, \frac{d}{4}+\frac{1}{4}\right) \geq \frac{12\left(1-\frac{3}{d}\right)^{\frac{d-3}{4}}\left(1+\frac{3}{d}\right)^{\frac{d+3}{4}}}{2^{\frac{d}{2}}(d+3)}.$$

The first summand on the left-hand side of this inequality can be expanded as

$$2B_{\frac{1}{2}+\frac{3}{2d}}\left(\frac{d}{4}-\frac{1}{4}, \frac{d}{4}+\frac{1}{4}\right) = 2B_{\frac{1}{2}}\left(\frac{d}{4}-\frac{1}{4}, \frac{d}{4}+\frac{1}{4}\right) + 2\int_{\frac{1}{2}}^{\frac{1}{2}+\frac{3}{2d}} x^{\frac{d}{4}-\frac{5}{4}}(1-x)^{\frac{d}{4}-\frac{3}{4}} dx.$$

As in Lemma 10.12,

$$(10.44) \quad 2B_{\frac{1}{2}}\left(\frac{d}{4}-\frac{1}{4}, \frac{d}{4}+\frac{1}{4}\right) - B\left(\frac{d}{4}-\frac{1}{4}, \frac{d}{4}+\frac{1}{4}\right) = \int_0^{1/2} x^{\frac{d}{4}-\frac{5}{4}}(1-x)^{\frac{d}{4}-\frac{5}{4}}(\sqrt{1-x}-\sqrt{x}) dx$$

can be bound below by

$$(10.45) \quad \frac{4}{2^{\frac{d}{2}}(d-1)}.$$

Similarly,  $x^{\frac{d}{4}-\frac{5}{4}}(1-x)^{\frac{d}{4}-\frac{3}{4}}$  is decreasing and concave on  $(\frac{1}{2}, \frac{1}{2}+\frac{3}{2d})$  for  $d \geq 5$ , so

$$(10.46) \quad 2\int_{\frac{1}{2}}^{\frac{1}{2}+\frac{3}{2d}} x^{\frac{d}{4}-\frac{5}{4}}(1-x)^{\frac{d}{4}-\frac{3}{4}} dx \geq \frac{6\left(\left(1-\frac{3}{d}\right)^{\frac{d-3}{4}}\left(1+\frac{3}{d}\right)^{\frac{d-5}{4}}+1\right)}{2^{\frac{d}{2}}d}.$$

Using the two lower bounds (10.45) and (10.46) in (10.43), it suffices to establish

$$\begin{aligned}
 (10.47) \quad \frac{4}{d-1} &\geq \frac{12 \left(1 - \frac{3}{d}\right)^{\frac{d-3}{4}} \left(1 + \frac{3}{d}\right)^{\frac{d+3}{4}}}{d+3} - \frac{6 \left( \left(1 - \frac{3}{d}\right)^{\frac{d-3}{4}} \left(1 + \frac{3}{d}\right)^{\frac{d-5}{4}} + 1 \right)}{d} \\
 &= 6 \left(1 - \frac{3}{d}\right)^{\frac{d-3}{4}} \left(1 + \frac{3}{d}\right)^{\frac{d-5}{4}} \left(2 \frac{d+3}{d^2} - \frac{1}{d}\right) - \frac{6}{d} \\
 &= 6 \left(1 - \frac{3}{d}\right)^{\frac{d-3}{4}} \left(1 + \frac{3}{d}\right)^{\frac{d-5}{4}} \frac{d+6}{d^2} - \frac{6}{d}
 \end{aligned}$$

The sequences

$$\left(1 - \frac{3}{d}\right)^{\frac{d+1}{4}}, \quad \left(1 + \frac{3}{d}\right)^{\frac{d-5}{4}}$$

are increasing, the product of their limits is 1. The inequality (10.47) is easy to verify (by hand or using a computer algebra system) for  $d = 1, 2, \dots, 16$ . Now assume  $d \in \mathbb{R}$  with  $d \geq 16$ . It is enough to prove

$$(10.48) \quad \frac{4}{d-1} + \frac{6}{d} \geq 6 \left(1 - \frac{3}{d}\right)^{-1} \frac{d+6}{d^2} = 6 \frac{d+6}{d(d-3)}.$$

Equivalently,

$$\frac{2(2d^2 - 33d + 27)}{(d-3)(d-1)d} \geq 0.$$

But this holds for all  $d \geq \frac{3}{4}(11 + \sqrt{97}) \approx 15.6366$ . ■

The proof of Theorem 10.1 is now complete.

## 11. BOUNDS ON THE MEDIAN AND THE EQUIPOINT OF THE BETA DISTRIBUTION

Like the median, the equipoint is a measure of central tendency in a probability distribution function (PDF). In this section we establish, for the Beta distribution, new lower bound for the median and, by relating the equipoint to the median, bounds on the equipoint needed in the proof of Theorem 1.2.

As in Section 1.8 we follow the convention that  $\mathfrak{s}, \mathfrak{t} \in \mathbb{R}_{>0}$ , and consider the Beta distribution  $\text{Beta}(\mathfrak{s}, \mathfrak{t})$  supported on  $[0, 1]$ . We denote by  $\varrho_{\mathfrak{s}, \mathfrak{t}}: [0, 1] \rightarrow \mathbb{R}$  the probability density function of  $\text{Beta}(\mathfrak{s}, \mathfrak{t})$ , i.e.,

$$\varrho_{\mathfrak{s}, \mathfrak{t}}(x) = \frac{x^{\mathfrak{s}-1}(1-x)^{\mathfrak{t}-1}}{B(\mathfrak{s}, \mathfrak{t})}$$

for  $x \in [0, 1]$ . The cumulative distribution function of  $\text{Beta}(\mathfrak{s}, \mathfrak{t})$  is  $I_x(\mathfrak{s}, \mathfrak{t})$  defined for  $x \in [0, 1]$ . We are interested in the **median**  $m_{\mathfrak{s}, \mathfrak{t}} \in [0, 1]$  of  $\text{Beta}(\mathfrak{s}, \mathfrak{t})$  and in the  **$(\mathfrak{s}, \mathfrak{t})$ -equipoint**  $e_{\mathfrak{s}, \mathfrak{t}} \in [0, 1]$  defined by

$$(11.1) \quad I_{m_{\mathfrak{s}, \mathfrak{t}}}(\mathfrak{s}, \mathfrak{t}) = \frac{1}{2}$$



and

$$(11.2) \quad I_{e_{\mathfrak{s},\mathfrak{t}}}(\mathfrak{s}, \mathfrak{t} + 1) + I_{e_{\mathfrak{s},\mathfrak{t}}}(\mathfrak{s} + 1, \mathfrak{t}) = 1$$

respectively. Here we used that  $I_x(\mathfrak{s}, \mathfrak{t})$  and  $I_x(\mathfrak{s}, \mathfrak{t} + 1) + I_x(\mathfrak{s} + 1, \mathfrak{t})$  are strictly monotonically increasing for  $x \in [0, 1]$ . We will continue to use this tacitly throughout this section.

**11.1. Lower bound for the equipoint  $e_{\mathfrak{s},\mathfrak{t}}$ .** In (11.2), if we move one of the two terms to the other side, we get the equivalent forms

$$I_{e_{\mathfrak{s},\mathfrak{t}}}(\mathfrak{s}, \mathfrak{t} + 1) = I_{1-e_{\mathfrak{s},\mathfrak{t}}}(\mathfrak{t}, \mathfrak{s} + 1), \quad I_{e_{\mathfrak{s},\mathfrak{t}}}(\mathfrak{s} + 1, \mathfrak{t}) = I_{1-e_{\mathfrak{s},\mathfrak{t}}}(\mathfrak{t} + 1, \mathfrak{s}).$$

**Lemma 11.1.** *For all  $\mathfrak{s}, \mathfrak{t} \in \mathbb{R}_{>0}$  and  $x \in [0, 1]$ , we have*

$$(a) \quad I_x(\mathfrak{s}, \mathfrak{t} + 1) + I_x(\mathfrak{s} + 1, \mathfrak{t}) = 2I_x(\mathfrak{s}, \mathfrak{t}) + (\mathfrak{s} - \mathfrak{t}) \frac{x^{\mathfrak{s}}(1-x)^{\mathfrak{t}}}{\mathfrak{s}\mathfrak{t}B(\mathfrak{s}, \mathfrak{t})}$$

$$(b) \quad I_x(\mathfrak{s}, \mathfrak{t} + 1) + I_x(\mathfrak{s} + 1, \mathfrak{t}) = 2I_x(\mathfrak{s} + 1, \mathfrak{t} + 1) + (1 - 2x) \frac{(\mathfrak{s} + \mathfrak{t})x^{\mathfrak{s}}(1-x)^{\mathfrak{t}}}{\mathfrak{s}\mathfrak{t}B(\mathfrak{s}, \mathfrak{t})}$$

*Proof.* Use the identities (8.17.20) and (8.17.21) from <http://dlmf.nist.gov/8.17#iv>. ■

Although there are some results on the median  $m_{\mathfrak{s},\mathfrak{t}}$  for special values of  $\mathfrak{s}, \mathfrak{t} \in \mathbb{R}_{\geq 0}$ <sup>4</sup>, about the only general thing that seems to be known [PYY89] is that

$$(11.3) \quad \mu_{\mathfrak{s},\mathfrak{t}} := \frac{\mathfrak{s}}{\mathfrak{s} + \mathfrak{t}} < m_{\mathfrak{s},\mathfrak{t}} < \frac{\mathfrak{s} - 1}{\mathfrak{s} + \mathfrak{t} - 2}$$

if  $1 < \mathfrak{t} < \mathfrak{s}$  (see also [Ker+] for an asymptotic analysis and numerical evidence in support of better bounds). The lower bound in (11.3) is actually the mean  $\mu_{\mathfrak{s},\mathfrak{t}}$  of  $\text{Beta}(\mathfrak{s}, \mathfrak{t})$  if  $\mathfrak{s}, \mathfrak{t} > 0$  and the upper bound is the mode of  $\text{Beta}(\mathfrak{s}, \mathfrak{t})$  if  $\mathfrak{s}, \mathfrak{t} > 1$ . In the next subsection we shall significantly improve the upper bound in (11.3).

Using Lemma 11.1(a), we see that

$$I_{m_{\mathfrak{s},\mathfrak{t}}}(\mathfrak{s}, \mathfrak{t} + 1) + I_{m_{\mathfrak{s},\mathfrak{t}}}(\mathfrak{s} + 1, \mathfrak{t}) = 2\frac{1}{2} + (\mathfrak{s} - \mathfrak{t}) \frac{m_{\mathfrak{s},\mathfrak{t}}^{\mathfrak{s}}(1 - m_{\mathfrak{s},\mathfrak{t}})^{\mathfrak{t}}}{\mathfrak{s}\mathfrak{t}B(\mathfrak{s}, \mathfrak{t})} \geq 1$$

and therefore

$$(11.4) \quad e_{\mathfrak{s},\mathfrak{t}} \leq m_{\mathfrak{s},\mathfrak{t}}$$

whenever  $\mathfrak{s}, \mathfrak{t} \in \mathbb{R}$  and  $0 < \mathfrak{t} \leq \mathfrak{s}$ . Using Lemma 11.1(b), we get

$$I_{m_{\mathfrak{s}+1,\mathfrak{t}+1}}(\mathfrak{s}, \mathfrak{t} + 1) + I_{m_{\mathfrak{s}+1,\mathfrak{t}+1}}(\mathfrak{s} + 1, \mathfrak{t}) =$$

$$2\frac{1}{2} + (1 - 2m_{\mathfrak{s}+1,\mathfrak{t}+1}) \frac{(\mathfrak{s} + \mathfrak{t})(m_{\mathfrak{s}+1,\mathfrak{t}+1})^{\mathfrak{s}}(1 - m_{\mathfrak{s}+1,\mathfrak{t}+1})^{\mathfrak{t}}}{\mathfrak{s}\mathfrak{t}B(\mathfrak{s}, \mathfrak{t})} \leq 1.$$

since  $m_{\mathfrak{s}+1,\mathfrak{t}+1} \geq \frac{\mathfrak{s} + 1}{\mathfrak{s} + \mathfrak{t} + 2} \geq \frac{\frac{\mathfrak{s}}{2} + \frac{\mathfrak{t}}{2} + 1}{\mathfrak{s} + \mathfrak{t} + 2} = \frac{1}{2}$  by (11.3). This shows that

$$(11.5) \quad m_{\mathfrak{s}+1,\mathfrak{t}+1} \leq e_{\mathfrak{s},\mathfrak{t}}$$

whenever  $\mathfrak{s}, \mathfrak{t} \in \mathbb{R}$  and  $0 < \mathfrak{t} < \mathfrak{s}$ .

---

<sup>4</sup>see [http://en.wikipedia.org/wiki/Beta\\_distribution](http://en.wikipedia.org/wiki/Beta_distribution)

These inequalities combine to give:

**Proposition 11.2.** *For  $\mathfrak{s}, \mathfrak{t} \in \mathbb{R}_{>0}$ ,*

$$(11.6) \quad e_{\mathfrak{s}, \mathfrak{t}} \leq m_{\mathfrak{s}, \mathfrak{t}} < \frac{\mathfrak{s} - 1}{\mathfrak{s} + \mathfrak{t} - 2}$$

when  $1 < \mathfrak{t} < \mathfrak{s}$  and

$$\frac{\mathfrak{s} + 1}{\mathfrak{s} + \mathfrak{t} + 2} < m_{\mathfrak{s}+1, \mathfrak{t}+1} \leq e_{\mathfrak{s}, \mathfrak{t}}$$

when  $0 < \mathfrak{t} < \mathfrak{s}$ . Later this lower bound on  $e_{\mathfrak{s}, \mathfrak{t}}$  proves important to us.

*Proof.* The first line of inequalities (11.6) follows from (11.3) and (11.4). The second from (11.3) and (11.5).  $\blacksquare$

**Remark 11.3.** The inequality (11.6) is easier to prove than the inequality  $e_{\mathfrak{s}, \mathfrak{t}} \leq \frac{\mathfrak{s}}{\mathfrak{s} + \mathfrak{t}}$  from Theorem 10.1; however, this weaker inequality seems not to be strong enough to prove Theorem 1.2.  $\square$

**11.2. New bounds on the median of the beta distribution.** Having a lower bound for the equipoint  $e_{\mathfrak{s}, \mathfrak{t}}$  in terms of the median  $m_{\mathfrak{s}+1, \mathfrak{t}+1}$ , we now turn our attention to the median of the beta distribution.

By Lemma 11.1(a), Simmons' inequality (10.2) is equivalent to

$$(11.7) \quad 2I_{\mu_{\mathfrak{s}, \mathfrak{t}}}(\mathfrak{s}, \mathfrak{t}) + (\mathfrak{s} - \mathfrak{t}) \frac{(\mu_{\mathfrak{s}, \mathfrak{t}})^{\mathfrak{s}} (1 - \mu_{\mathfrak{s}, \mathfrak{t}})^{\mathfrak{t}}}{\mathfrak{s} \mathfrak{t} B(\mathfrak{s}, \mathfrak{t})} \geq 1$$

which is therefore conjectured for all  $\mathfrak{s}, \mathfrak{t} \in \mathbb{R}$  with  $0 < \mathfrak{s} \leq \mathfrak{t}$ . Proposition 11.5 below proves a weakening of (11.7) where an extra factor of 2 is introduced in the second term on the left hand side.

**Lemma 11.4.** *Suppose  $\mathfrak{s}, \mathfrak{t} \in \mathbb{R}_{>0}$  and set  $\mu := \mu_{\mathfrak{s}, \mathfrak{t}}$ . Then*

$$(11.8) \quad \int_0^\mu (\mu - x)^{\mathfrak{s}-1} (1 - \mu + x)^{\mathfrak{t}-1} x \, dx = \int_0^{1-\mu} (\mu + x)^{\mathfrak{s}-1} (1 - \mu - x)^{\mathfrak{t}-1} x \, dx$$

$$= \frac{\mu^{\mathfrak{s}} (1 - \mu)^{\mathfrak{t}}}{\mathfrak{s} + \mathfrak{t}} = \frac{\mathfrak{s}^{\mathfrak{s}} \mathfrak{t}^{\mathfrak{t}}}{(\mathfrak{s} + \mathfrak{t})^{\mathfrak{s} + \mathfrak{t} + 1}}.$$

*Proof.* Reversing the direction of integration in the first integral and changing the domain of integration in the second integral, we get

$$(11.9) \quad \int_0^\mu (\mu - x)^{\mathfrak{s}-1} (1 - \mu + x)^{\mathfrak{t}-1} x \, dx = \int_0^\mu x^{\mathfrak{s}-1} (1 - x)^{\mathfrak{t}-1} (\mu - x) dx,$$

$$(11.10) \quad \int_0^{1-\mu} (\mu + x)^{\mathfrak{s}-1} (1 - \mu - x)^{\mathfrak{t}-1} x \, dx = \int_\mu^1 x^{\mathfrak{s}-1} (1 - x)^{\mathfrak{t}-1} (x - \mu) dx.$$

If we subtract (11.9) from (11.10) and divide by  $B(\mathfrak{s}, \mathfrak{t})$ , we get

$$\int_0^1 \varrho_{\mathfrak{s}, \mathfrak{t}}(x) (x - \mu) dx = \mu - \mu = 0$$

by the definition of the mean  $\mu$ . So the first equality is proved. On the other hand, if we add (11.9) and (11.10) and divide again by  $B(\mathfrak{s}, \mathfrak{t})$ , we get

$$\int_0^1 \varrho_{\mathfrak{s}, \mathfrak{t}}(x) |x - \mu| dx$$

which is by the formula for the *mean absolute deviation* of  $\text{Beta}(\mathfrak{s}, \mathfrak{t})$  (cf. the proof of [DZ91, Corollary 1]) equal to

$$\frac{2\mu(1-\mu)}{\mathfrak{s} + \mathfrak{t}} \varrho_{\mathfrak{s}, \mathfrak{t}}(\mu),$$

thus showing the second equation. The third equation in (11.8) is clear.  $\blacksquare$

**Proposition 11.5.** *Suppose  $1 \leq \mathfrak{t} \leq \mathfrak{s}$  such that  $\mathfrak{s} + \mathfrak{t} \geq 3$  and set  $\mu := \mu_{\mathfrak{s}, \mathfrak{t}}$ . Then we have*

$$2I_\mu(\mathfrak{s}, \mathfrak{t}) + 2(\mathfrak{s} - \mathfrak{t}) \frac{\mu^{\mathfrak{s}}(1-\mu)^{\mathfrak{t}}}{\mathfrak{s}\mathfrak{t}B(\mathfrak{s}, \mathfrak{t})} \geq 1$$

*Proof.* We have to show

$$I_\mu(\mathfrak{s}, \mathfrak{t}) + 2(\mathfrak{s} - \mathfrak{t}) \frac{\mu^{\mathfrak{s}}(1-\mu)^{\mathfrak{t}}}{\mathfrak{s}\mathfrak{t}B(\mathfrak{s}, \mathfrak{t})} \geq I_{1-\mu}(\mathfrak{t}, \mathfrak{s})$$

which is equivalent to

$$B_\mu(\mathfrak{s}, \mathfrak{t}) + 2\chi \geq B_{1-\mu}(\mathfrak{t}, \mathfrak{s})$$

where

$$\chi := \frac{\mathfrak{s} - \mathfrak{t}}{\mathfrak{s}\mathfrak{t}} \mu^{\mathfrak{s}}(1-\mu)^{\mathfrak{t}}.$$

This means

$$2\chi + \int_0^\mu x^{\mathfrak{s}-1}(1-x)^{\mathfrak{t}-1} dx \geq \int_0^{1-\mu} x^{\mathfrak{t}-1}(1-x)^{\mathfrak{s}-1} dx$$

which we rewrite as

$$\chi + \int_0^\mu (\mu - x)^{\mathfrak{s}-1}(1-\mu+x)^{\mathfrak{t}-1} dx \geq -\chi + \int_0^{1-\mu} (\mu+x)^{\mathfrak{s}-1}(1-\mu-x)^{\mathfrak{t}-1} dx.$$

We have  $\frac{1}{2} \leq \mu \leq 1$  and therefore  $0 \leq 1-\mu \leq \frac{1}{2} \leq \mu \leq 1$ . In particular, the domain of integration is smaller on the left hand side. The idea is to compare the two terms under the integral pointwise on  $[0, 1-\mu]$  after correcting these two terms using  $\chi$  and  $-\chi$ , respectively. The two terms agree when substituting  $x = 0$ . The derivative at  $x = 0$  of the term under the integral on the left hand side is by the product rule the negative term

$$\mu^{\mathfrak{s}-2}(1-\mu)^{\mathfrak{t}-2}(\mu(\mathfrak{s} + \mathfrak{t} - 2) - \mathfrak{s} + 1) = \mu^{\mathfrak{s}-2}(1-\mu)^{\mathfrak{t}-2} \frac{\mathfrak{t} - \mathfrak{s}}{\mathfrak{s} + \mathfrak{t}}$$

and on the right hand side it is the additive inverse. We want to counterbalance the derivatives at  $x = 0$  by adding and subtracting a multiple of the term from Lemma 11.4 on the left and right hand side, respectively. The derivative of that latter term at  $x = 0$  is of course

$$\mu^{\mathfrak{s}-1}(1-\mu)^{\mathfrak{t}-1} = \mu^{\mathfrak{s}-2}(1-\mu)^{\mathfrak{t}-2} \frac{\mathfrak{s}\mathfrak{t}}{(\mathfrak{s} + \mathfrak{t})^2}.$$

We thus would like to add

$$c := \frac{(\mathfrak{s} - \mathfrak{t})(\mathfrak{s} + \mathfrak{t})}{\mathfrak{s}\mathfrak{t}} = \frac{\mathfrak{s}^2 - \mathfrak{t}^2}{\mathfrak{s}\mathfrak{t}}$$

times the term from Lemma 11.4 on the left hand side and subtract it on the right hand side. The miracle now is that this is exactly  $\chi$ . Our claim can thus be rewritten as

$$\int_0^\mu (\mu - x)^{s-1} (1 - \mu + x)^{t-1} (1 + cx) dx \geq \int_0^{1-\mu} (\mu + x)^{s-1} (1 - \mu - x)^{t-1} (1 - cx) dx.$$

The two terms under the integral now take the same value at  $x = 0$  and have the same derivative there. There is now a hope to show for  $x \in [0, 1 - \mu]$  that the term on the left hand side is pointwise less than or equal the term on the right hand side. We will do this and thus even show the stronger claim that

$$\int_0^{1-\mu} (\mu - x)^{s-1} (1 - \mu + x)^{t-1} (1 + cx) dx \geq \int_0^{1-\mu} (\mu + x)^{s-1} (1 - \mu - x)^{t-1} (1 - cx) dx.$$

If we define (noting that  $1 - cx \geq 1 - c(1 - \mu) = 1 - c\frac{t}{s+t} = 1 - \frac{s-t}{s} = \frac{t}{s} > 0$ )

$$g: [0, 1 - \mu] \rightarrow \mathbb{R}, \quad x \mapsto \left( \frac{\mu - x}{\mu + x} \right)^{s-1} \left( \frac{1 - \mu + x}{1 - \mu - x} \right)^{t-1} \frac{1 + cx}{1 - cx},$$

it is thus enough to show that  $g(x) \geq 1$  for all  $x \in [0, 1 - \mu]$ . Clearly we have  $g(0) = 1$ . So it is enough to show that  $g'(x) \geq 0$  for all  $x \in [0, 1 - \mu]$ . A straightforward calculation shows

$$g'(x) = \frac{2 \left( \frac{\mu - x}{\mu + x} \right)^s \left( \frac{1 - \mu + x}{1 - \mu - x} \right)^t}{(1 - cx)^2 (1 - \mu + x)^2 (\mu - x)^2} + x^2 h(x^2)$$

where

$$h: \begin{cases} [0, (1 - \mu)^2] \rightarrow \mathbb{R} \\ y \mapsto \frac{(s-t)(s+t-3)}{s+t} - \frac{(s+t)(s^4 - s^3(2t+1) + 2s^2t + 2s(t-1)t^2 - (t-1)t^3)}{s^2t^2} y. \end{cases}$$

Since  $h$  is linear, it is thus enough to show that  $h(0) \geq 0$  and  $h((1 - \mu)^2) \geq 0$ . The first condition follows from the hypothesis  $s + t \geq 3$ . Another straightforward calculation shows

$$h((1 - \mu)^2) = \frac{(t-1)(2s^2 - 3st + t^2)}{s^2}.$$

Because of  $t \geq 1$  it remains only to show that  $2s^2 - 3st + t^2 \geq 0$ . Now we have

$$2s^2 - 3st + t^2 = (s-t)(2s-t) \geq 0$$

since  $s \geq t$ . ■

The following corollary improves the previously known upper bound (11.3) on the median  $m_{s,t}$  in the case where  $1 < t \leq s$  and  $s + t \geq 3$  because of the following lemma.

**Lemma 11.6.** *Suppose  $s, t \in \mathbb{R}$  such that  $s \geq t \geq 1$  and  $s + t > 2$ . Then*

$$\frac{s}{s+t} + \frac{s-t}{(s+t)^2} \leq \frac{s-1}{s+t-2}.$$

*Proof.* A straightforward calculation yields

$$\frac{s-1}{s+t-2} - \frac{s}{s+t} - \frac{s-t}{(s+t)^2} = \frac{2(s-t)}{(s+t-2)(s+t)^2} \geq 0. \quad \blacksquare$$

**Corollary 11.7.** *Suppose  $1 \leq \mathfrak{t} \leq \mathfrak{s}$  such that  $\mathfrak{s} + \mathfrak{t} \geq 3$ . Then we have*

$$\mu_{\mathfrak{s},\mathfrak{t}} = \frac{\mathfrak{s}}{\mathfrak{s} + \mathfrak{t}} \leq m_{\mathfrak{s},\mathfrak{t}} \leq \mu_{\mathfrak{s},\mathfrak{t}} + \frac{\mathfrak{s} - \mathfrak{t}}{(\mathfrak{s} + \mathfrak{t})^2}.$$

*Proof.* The first inequality comes from [PYY89]. To prove the second, we have to show that

$$2I_{\mu}(\mathfrak{s}, \mathfrak{t}) + 2 \int_{\mu}^{\mu + \frac{\mathfrak{s}-\mathfrak{t}}{(\mathfrak{s}+\mathfrak{t})^2}} \varrho_{\mathfrak{s},\mathfrak{t}}(x) dx \geq 1$$

where  $\mu := \mu_{\mathfrak{s},\mathfrak{t}}$ . By Proposition 11.5, it is henceforth enough to show that

$$\int_{\mu}^{\mu + \frac{\mathfrak{s}-\mathfrak{t}}{(\mathfrak{s}+\mathfrak{t})^2}} \varrho_{\mathfrak{s},\mathfrak{t}}(x) dx \geq (\mathfrak{s} - \mathfrak{t}) \frac{\mu^{\mathfrak{s}}(1 - \mu)^{\mathfrak{t}}}{\mathfrak{s}\mathfrak{t}B(\mathfrak{s}, \mathfrak{t})}.$$

This is trivial if  $\mathfrak{s} = \mathfrak{t}$ . If  $1 < \mathfrak{t}$  (and therefore  $1 < \mathfrak{s}$ ), then  $\frac{\mathfrak{s}-1}{\mathfrak{s}+\mathfrak{t}-2}$  is the mode of  $\text{Beta}(\mathfrak{s}, \mathfrak{t})$  and by Lemma 11.6 we have

$$\varrho_{\mathfrak{s},\mathfrak{t}}(x) \geq \varrho_{\mathfrak{s},\mathfrak{t}}(\mu)$$

for all  $x \in [\mu, \mu + \frac{\mathfrak{s}-\mathfrak{t}}{(\mathfrak{s}+\mathfrak{t})^2}]$ . Therefore it is enough to show that

$$\frac{\mathfrak{s} - \mathfrak{t}}{(\mathfrak{s} + \mathfrak{t})^2} \varrho_{\mathfrak{s},\mathfrak{t}}(\mu) \geq (\mathfrak{s} - \mathfrak{t}) \frac{\mu^{\mathfrak{s}}(1 - \mu)^{\mathfrak{t}}}{\mathfrak{s}\mathfrak{t}B(\mathfrak{s}, \mathfrak{t})}$$

but this holds even with equality since

$$\frac{1}{(\mathfrak{s} + \mathfrak{t})^2} = \frac{\mu(1 - \mu)}{\mathfrak{s}\mathfrak{t}}. \quad \blacksquare$$

The following table illustrates the quality of the lower bound  $\mu_{\mathfrak{s},\mathfrak{t}}$  on the median  $m_{\mathfrak{s},\mathfrak{t}}$  (for  $1 \leq \mathfrak{t} \leq \mathfrak{s}$ ) and the quality of the new upper bound  $\mu_{\mathfrak{s},\mathfrak{t}} + \frac{\mathfrak{s}-\mathfrak{t}}{(\mathfrak{s}+\mathfrak{t})^2}$  (for  $1 \leq \mathfrak{t} \leq \mathfrak{s}$  with  $\mathfrak{s} + \mathfrak{t} \geq 3$ ) as compared to the less tight old upper bound  $\frac{\mathfrak{s}-1}{\mathfrak{s}+\mathfrak{t}-2}$ . If one assumes that (11.7) is true for all real  $\mathfrak{s}, \mathfrak{t}$  (as opposed to  $\mathfrak{s}, \mathfrak{t} \in \frac{1}{2}\mathbb{N}$  as given by Theorem 10.1) with  $1 \leq \mathfrak{t} \leq \mathfrak{s}$  with  $\mathfrak{s} + \mathfrak{t} \geq 3$ , then one can deduce along the lines of Corollary 11.7 an even better upper bound on  $m_{\mathfrak{s},\mathfrak{t}}$  for  $1 \leq \mathfrak{t} \leq \mathfrak{s}$  with  $\mathfrak{s} + \mathfrak{t} \geq 3$ , namely  $\mu_{\mathfrak{s},\mathfrak{t}} + \frac{\mathfrak{s}-\mathfrak{t}}{2(\mathfrak{s}+\mathfrak{t})^2}$  which we therefore also include in the table.

$\mathfrak{s}$	$\mathfrak{t}$	$\mu_{\mathfrak{s},\mathfrak{t}}$	$m_{\mathfrak{s},\mathfrak{t}}$	$\mu_{\mathfrak{s},\mathfrak{t}} + \frac{\mathfrak{s}-\mathfrak{t}}{2(\mathfrak{s}+\mathfrak{t})^2}$	$\mu_{\mathfrak{s},\mathfrak{t}} + \frac{\mathfrak{s}-\mathfrak{t}}{(\mathfrak{s}+\mathfrak{t})^2}$	$\frac{\mathfrak{s}-1}{\mathfrak{s}+\mathfrak{t}-2}$
2.5	1	0.714286	0.757858	0.77551	0.836735	1
3	1	0.75	0.793701	0.8125	0.875	1
3	2	0.6	0.614272	0.62	0.64	0.666667
4	2	0.666667	0.68619	0.694444	0.722222	0.75
10	3	0.769231	0.783314	0.789941	0.810651	0.818182
10	7	0.588235	0.591773	0.593426	0.598616	0.6

## 12. PROOF OF THEOREM 1.2

In this section we prove Theorem 1.2 by establishing Proposition 9.2. We start by tweaking Problem 9.1:

**Problem 12.1.** Given a positive integer  $d$ , minimize

$$f_{s,t}(\sigma) = \frac{2(1-\sigma)sI_{1-\sigma}\left(\frac{t}{2}, 1 + \frac{s}{2}\right) + 2\sigma tI_{\sigma}\left(\frac{s}{2}, 1 + \frac{t}{2}\right)}{(1-\sigma)s + \sigma t} - 1$$

subject to the constraints

- (i)  $s, t \in \mathbb{N}$  and  $s + t = d$ ;
- (ii)  $s \geq \frac{d}{2}$ ;
- (iii)  $0 \leq \sigma \leq \frac{s}{d}$ ; and
- (iv)  $I_{\sigma}\left(\frac{s}{2}, \frac{t}{2} + 1\right) = I_{1-\sigma}\left(\frac{t}{2}, \frac{s}{2} + 1\right)$ .

Problem 12.1 is equivalent to Problem 9.1. Indeed, the only difference is the interval for  $\sigma$  in (iii). However, by Section 10 we know that  $\sigma_{s,t} = \sigma \in [0, 1]$ , the solution to (iv) will automatically satisfy  $\sigma_{s,t} \leq \frac{s}{d}$ .

**12.1. An auxiliary function.** For  $s, t \in \mathbb{R}_{>0}$  let

$$(12.1) \quad g_{s,t}(\sigma) := -1 + I_{\sigma}\left(\frac{s}{2}, \frac{t}{2} + 1\right) + I_{1-\sigma}\left(\frac{t}{2}, \frac{s}{2} + 1\right).$$

**Lemma 12.2.** For  $s, t \in \mathbb{R}_{>0}$ , we have

$$f_{s,t}(\sigma_{s,t}) = g_{s,t}(\sigma_{s,t}) = 2 I_{\sigma_{s,t}}\left(\frac{s}{2}, 1 + \frac{t}{2}\right) - 1.$$

Thus at the minimizer of  $f_{s,t}$ , the functions  $f_{s,t}$  and  $g_{s,t}$  have the same value.

*Proof.* This is straightforward since  $f_{s,t}$  assumes its minimum where the two incomplete beta expressions (appearing in both  $f_{s,t}$  and  $g_{s,t}$ ) are equal. ■

**Lemma 12.3.** The function  $g_{s,t}$  can be rewritten as

$$(12.2) \quad g_{s,t}(\sigma) = \sigma^{s/2}(1-\sigma)^{t/2} \frac{\Gamma\left(\frac{s}{2} + \frac{t}{2} + 1\right)}{\Gamma\left(\frac{s}{2} + 1\right) \Gamma\left(\frac{t}{2} + 1\right)}.$$

*Proof.* First recall that  $I_{1-x}(b, a) = 1 - I_x(a, b)$  and apply this to the second incomplete beta summand in the definition of  $g_{s,t}$ :

$$g_{s,t}(\sigma) = I_{\sigma}\left(\frac{s}{2}, \frac{t}{2} + 1\right) - I_{\sigma}\left(\frac{s}{2} + 1, \frac{t}{2}\right).$$

Now use recursive formulas for  $I_{\sigma}$ <sup>5</sup> and simplify. ■

---

<sup>5</sup>Equations (8.17.20) and (8.17.21) in <http://dlmf.nist.gov/8.17>.

**Lemma 12.4.** *The function  $g_{s,t}$  is monotonically increasing on  $\left[0, \frac{s}{s+t}\right]$  whenever  $s, t \in \mathbb{R}_{>0}$  with  $s \geq t$ .*

*Proof.* Using Lemma 12.3 it is easy to see that

$$g'_{s,t}(\sigma) = -\frac{1}{2}\sigma^{\frac{s}{2}-1}(1-\sigma)^{\frac{t}{2}-1}(s(\sigma-1)+\sigma t)\frac{\Gamma\left(\frac{s}{2}+\frac{t}{2}+1\right)}{\Gamma\left(\frac{s}{2}+1\right)\Gamma\left(\frac{t}{2}+1\right)}. \quad \blacksquare$$

We shall exploit bounds on  $\sigma_{s,t}$ . The lower bound can be deduced from our results in Section 11:

**Corollary 12.5.** *For  $s, t \in \mathbb{N}$  with  $s \geq t$  we have*

$$(12.3) \quad \sigma_{s,t} \geq \frac{s+2}{s+t+4}.$$

*Proof.* Simply note that in the notation of Section 11,  $\sigma_{s,t} = e_{\frac{s}{2}, \frac{t}{2}}$  and use Proposition 11.2.  $\blacksquare$

Combining this lower bound for  $\sigma_{s,t}$  with Theorem 10.1, we have for  $s, t \in \mathbb{N}$  with  $s \geq t$ ,

$$(12.4) \quad \phi(s, t) := \frac{s+2}{s+t+4} \leq \sigma_{s,t} \leq \frac{s}{s+t} =: \psi(s, t).$$

**Lemma 12.6.** *For  $s, t \in \mathbb{N}$  with  $s \geq t$  we have*

$$(12.5) \quad g_{s,t}(\phi(s, t)) \leq g_{s,t}(\sigma_{s,t}) = f_{s,t}(\sigma_{s,t}) = g_{s,t}(\sigma_{s,t}) \leq g_{s,t}(\psi(s, t)).$$

*Proof.* This follows from the monotonicity of  $g_{s,t}$  on  $\left[0, \frac{s}{s+t}\right]$  and by the coincidence of  $f_{s,t}$  and  $g_{s,t}$  in the equipoint  $\sigma_{s,t}$ .  $\blacksquare$

**12.2. Two step monotonicity of  $f_{s,t}(\sigma_{s,t})$ .** In this subsection, in Proposition 12.8, we show for  $s, t \in \mathbb{N}$  with  $s \geq t$  that  $f_{s,t}(\sigma_{s,t}) \leq f_{s+2,t-2}(\sigma_{s+2,t-2})$ .

**Lemma 12.7.** *If  $s, t \in \mathbb{R}_{>0}$  with  $s \geq t$  then*

$$(12.6) \quad g_{s+2,t-2}(\phi(s+2, t-2)) \geq g_{s,t}(\psi(s, t)).$$

*Proof.* With  $d = s+t$ , (12.6) is equivalent to

$$(12.7) \quad (4+s)^{s+2}d^d \geq s^s(4+d)^d(2+s)^2.$$

This follows from (12.2) and the identities  $B(\alpha+1, \beta) = \frac{\alpha}{\alpha+\beta}B(\alpha, \beta)$  and  $B(\alpha, \beta+1) = \frac{\beta}{\alpha+\beta}B(\alpha, \beta)$ .

Rewrite (12.7) into

$$(12.8) \quad \left(1 + \frac{4}{d}\right)^d \leq \left(1 + \frac{4}{s}\right)^s \left(\frac{s+4}{s+2}\right)^2 =: \xi(s).$$

We claim the right-hand side  $\xi(s)$  is an increasing function of  $s$  on  $\mathbb{R}_{\geq 0}$ . Indeed, using  $s = 2S$ ,

$$\begin{aligned}\Xi(S) &= \xi\left(\frac{s}{2}\right) = \left(1 + \frac{2}{S}\right)^{2S} \left(\frac{S+2}{S+1}\right)^2 \\ &= \left(1 + \frac{2}{S}\right)^S \cdot \left(1 + \frac{2}{S}\right)^S \left(\frac{S+2}{S+1}\right)^2.\end{aligned}$$

The first factor in the last expression is well-known to be an increasing function with limit  $e^2$ . Let

$$\zeta(S) := \left(1 + \frac{2}{S}\right)^S \left(\frac{S+2}{S+1}\right)^2.$$

Then

$$\zeta'(S) = \frac{(S+2)^2 \left(\frac{S+2}{S}\right)^S ((S+1) \log\left(\frac{S+2}{S}\right) - 2)}{(S+1)^3}.$$

Observe that  $(S+1) \log\left(\frac{S+2}{S}\right) - 2 \geq 0$  iff

$$\left(1 + \frac{2}{S}\right)^{S+1} \geq e^2.$$

But it is well-known and easy to see that this left-hand side is decreasing with limit  $e^2$ . This shows that  $\zeta'(S) \geq 0$  and hence the right-hand side of (12.8) is an increasing function of  $s$ .

It thus suffices to show

$$(12.9) \quad \left(\frac{d+4}{d}\right)^d \leq \xi\left(\frac{d}{2}\right) = \left(\frac{d+8}{d+4}\right)^2 \left(\frac{d+8}{d}\right)^{\frac{d}{2}},$$

or equivalently,

$$(12.10) \quad \left(\frac{d+4}{d}\right)^{\frac{d}{2}} \leq \left(\frac{d+8}{d+4}\right)^{2+\frac{d}{2}}.$$

Again, we show this hold for  $d \in \mathbb{R}_{>0}$ . Writing  $d = 4D$ , (12.10) becomes

$$(12.11) \quad \left(\frac{D+1}{D}\right)^{2D} \leq \left(\frac{D+2}{D+1}\right)^{2+2D}.$$

So it suffices to establish

$$\left(1 + \frac{1}{D}\right)^D \leq \left(1 + \frac{1}{D+1}\right)^{D+1}.$$

But this is well-known or easy to establish using calculus. ■

**Proposition 12.8.** *For  $s, t \in \mathbb{N}$  with  $s \geq t$  we have*

$$(12.12) \quad f_{s,t}(\sigma_{s,t}) \leq f_{s+2,t-2}(\sigma_{s+2,t-2}).$$



*Proof.* Observe that

$$\begin{aligned}
 f_{s+2,t-2}(\sigma_{s+2,t-2}) &\stackrel{(12.5)}{\geq} g_{s+2,t-2}(\phi(s+2, t-2)) \\
 &\stackrel{(12.6)}{\geq} g_{s,t}(\psi(s, t)) \\
 &\stackrel{(12.5)}{\geq} f_{s,t}(\sigma_{s,t}). \quad \blacksquare
 \end{aligned}$$

**12.3. Boundary cases.** Having established for each  $d$  monotonicity in  $s$  of  $f_{s,d-s}(\sigma_{s,d-s})$ , where  $\frac{d}{2} \leq s \leq d-2$ , we now turn to the boundary cases.

**Lemma 12.9.** *For  $s \in \mathbb{N}$  we have*

$$(12.13) \quad g_{s+1,s-1}(\phi(s+1, s-1)) \geq g_{s,s}(\psi(s, s)).$$

**Remark 12.10.** The proof uses Chu's inequality (see e.g. [MV70, p. 288]) on the quotient of gamma functions, which says that for  $s \in \mathbb{N}$ ,

$$\sqrt{\frac{2s+1}{4}} \leq \frac{\Gamma(\frac{s}{2}+1)}{\Gamma(\frac{s+1}{2})} \leq \frac{s+1}{\sqrt{2s+1}}.$$

Thus, it is at this point the assumption that  $s$  is an integer is used.  $\square$

*Proof.* Inequality (12.13) is equivalent to

$$(12.14) \quad \frac{(s+2)^{-s}((s+1)(s+3))^{\frac{s+1}{2}} \Gamma(\frac{s}{2}+1)^2}{2\Gamma(\frac{s+3}{2})^2} \geq 1.$$

Further, using Chu's inequality it suffices to establish

$$(12.15) \quad \left(1 - \frac{1}{(s+2)^2}\right)^{\frac{s+1}{2}} \geq 1 - \frac{1}{2s+5}.$$

Equivalently,

$$(12.16) \quad \left(1 - \frac{1}{(s+2)^2}\right)^{\frac{s+1}{2}(2s+5)} \geq \left(1 - \frac{1}{2s+5}\right)^{2s+5}.$$

The right-hand side of (12.16) is increasing with limit as  $s \rightarrow \infty$  being  $e^{-1}$ . Further, as

$$\frac{s+1}{2}(2s+5) \leq (s+2)^2 - 1 \quad \text{for } s \geq -1,$$

we have

$$(12.17) \quad \left(1 - \frac{1}{(s+2)^2}\right)^{\frac{s+1}{2}(2s+5)} \geq \left(1 - \frac{1}{(s+2)^2}\right)^{(s+2)^2-1}$$

Now consider

$$\zeta(x) := \left(1 - \frac{1}{x^2}\right)^{x^2-1}.$$

We claim it is (for  $x > 1$ ) decreasing. Indeed,

$$\zeta'(x) = \frac{2 \left(1 - \frac{1}{x^2}\right)^{x^2} x \left(x^2 \log \left(1 - \frac{1}{x^2}\right) + 1\right)}{x^2 - 1}$$

and

$$x^2 \log \left(1 - \frac{1}{x^2}\right) + 1 < 0$$

since

$$\left(1 - \frac{1}{x^2}\right)^{x^2}$$

is increasing with limit  $e^{-1}$ .

Now the left-hand side of (12.16) is greater than the right-hand side of (12.17) which is decreasing with  $s$  towards  $e^{-1}$  which is an upper bound on the right-hand side of (12.16). ■

**Lemma 12.11.** *For  $s \in \mathbb{R}$  with  $s \geq 1$  we have*

$$(12.18) \quad g_{s+2,s-1}(\phi(s+2, s-1)) \geq g_{s+1,s}(\psi(s+1, s)).$$

*Proof.* Expanding  $g$ 's as was done to obtain (12.7), we see (12.18) is equivalent to

$$(12.19) \quad \xi(s) := \frac{s+4}{s+2} \left(1 + \frac{4}{s}\right)^{\frac{s}{2}} \left(\frac{2s+1}{2s+5}\right)^{s+\frac{1}{2}} \geq 1$$

Letting  $s = 2S$ , consider

$$\Xi(S) = \xi\left(\frac{s}{2}\right) = \frac{S+2}{S+1} \left(1 + \frac{2}{S}\right)^S \left(1 - \frac{4}{4S+5}\right)^{2S+\frac{1}{2}}.$$

We have to show that  $\Xi(S) \geq 1$  for  $S \in [\frac{1}{2}, \infty)$ . To this end, we will show that  $\Xi'(S) \leq 0$  for  $S \in [\frac{1}{2}, \infty)$  and  $\lim_{S \rightarrow \infty} \Xi(S) = 1$ . For the latter, note that

$$\lim_{S \rightarrow \infty} \left(1 - \frac{4}{4S+5}\right)^{2S+\frac{1}{2}} = \lim_{S \rightarrow \infty} \left(1 - \frac{4}{4S+5}\right)^{-2} \sqrt{\lim_{S \rightarrow \infty} \left(1 - \frac{4}{4S+5}\right)^{5+4S}} = e^{-4}$$

and therefore

$$\lim_{S \rightarrow \infty} \Xi(S) = e^2 \sqrt{e^{-4}} = 1.$$

A straightforward computation shows

$$\Xi'(S) = \frac{\left(1 + \frac{2}{S}\right)^S \left(1 - \frac{4}{4S+5}\right)^{2S+\frac{1}{2}}}{(S+1)^2(4S+5)} \Xi_1(S)$$

where

$$\begin{aligned} \Xi_1(S) &:= 1 + 2S + (S+1)(S+2)(4S+5) \\ &\quad \log \left(1 - \frac{8}{4S+5} + \frac{16}{(4S+5)^2} - \frac{16}{(4S+5)S} + \frac{32}{(4S+5)^2 S} + \frac{2}{S}\right). \end{aligned}$$

It is enough to show that  $\Xi_1(S) \leq 0$  for  $S \in [\frac{1}{2}, \infty)$ . Using  $\log x \leq x - 1$  for  $x > 0$ , we have

$$\begin{aligned} \Xi_1(S) &\leq \\ 1 + 2S + (S+1)(S+2)(4S+5) &\left( -\frac{8}{4S+5} + \frac{16}{(4S+5)^2} - \frac{16}{(4S+5)S} + \frac{32}{(4S+5)^2S} + \frac{2}{S} \right) \\ &= \frac{9}{5(4S+5)} + \frac{4}{5S} - 2 \end{aligned}$$

which evaluates to  $-\frac{1}{7}$  for  $S = \frac{1}{2}$  and therefore is clearly negative for  $S \geq \frac{1}{2}$ .  $\blacksquare$

*Proof of Proposition 9.2.* Suppose  $d = s + t$  is an even integer. If  $s$  and  $t$  are even integers with  $s \geq t$ , then two step monotonicity in Proposition 12.8 tells us that the minimizer over  $r$  even with  $s - r \geq t + r$ , of  $f_{s-r, t+r}(\sigma_{s-r, t+r})$  occurs where  $s - r = t + r$ ; that is,  $s = \frac{d}{2} = t$ . If  $s$  and  $t$  are odd integers with  $s \geq t$ , then stepping  $r$  by 2 preserves odd integers, so two step monotonicity gives that the minimizer in this case is at  $s = \frac{d}{2} + 1$  and  $t = \frac{d}{2} - 1$ . Compare the two minimizers using Lemmas 12.7 and 12.9 which give:

$$(12.20) \quad f_{s,s}(\sigma_{s,s}) \leq g_{s,s}(\psi(s, s)) \leq g_{s+1, s-1}(\phi(s+1, s-1)) \leq f_{s+1, s-1}(\sigma_{s+1, s-1}).$$

Apply this inequality to  $s = \frac{d}{2}$  to get the minimizer is  $\frac{d}{2}$ .

Suppose  $d$  is an odd integer. As before, Proposition 12.8 gives that the minimizer of  $f_{s,t}(\sigma_{s,t})$  over  $s$  odd and  $t$  even is  $s = \frac{d}{2} + \frac{1}{2}$  and  $t = \frac{d}{2} - \frac{1}{2}$ . Likewise minimizing over  $t$  odd and  $s$  even yields a minimizer which compares to the previous one unfavorably (using Lemmas 12.7 and 12.11).  $\blacksquare$

### 13. ESTIMATING $\vartheta(d)$ FOR ODD $d$ .

Recall that  $\vartheta(d)$  ( $d \in \mathbb{N}$ ) has been introduced in (1.1) and was simplified in Proposition 4.2. It was explicitly determined in (1.2) for even  $d$  by the expression which we repeat in part (b) of Theorem 13.1. For odd  $d$ , we have only the implicit characterization of Theorem 1.2. We do not know a way of making this more explicit. In this section, we exhibit however, for odd  $d$ , a compact interval containing  $\vartheta(d)$  whose end-points are given by nice analytic expressions in  $d$ . For the upper end point of the interval, we provide two versions: one involving the gamma function only and another which seems to be even tighter but involves the regularized beta function. The main result of this section is

**Theorem 13.1.** *Let  $d \in \mathbb{N}$ .*

- (a)  $\vartheta(1) = 1$
- (b) *Suppose  $d$  is even. Then*

$$\vartheta(d) = \sqrt{\pi} \frac{\Gamma(1 + \frac{d}{4})}{\Gamma(\frac{1}{2} + \frac{d}{4})}.$$

- (c) *Suppose  $d \geq 3$  is odd. Then there is a unique  $p \in [0, 1]$  satisfying*

$$I_p\left(\frac{d+1}{4}, \frac{d+3}{4}\right) = I_{1-p}\left(\frac{d-1}{4}, \frac{d+5}{4}\right).$$

For this  $p$ , we have  $p \in [\frac{1}{2}, \frac{d+1}{2d}]$ ,

$$(13.1) \quad \vartheta_-(d) \leq \vartheta(d) = \frac{\Gamma(\frac{d+3}{4}) \Gamma(\frac{d+5}{4})}{p^{\frac{d-1}{4}} (1-p)^{\frac{d+1}{4}} \Gamma(\frac{d}{2} + 1)} \leq \min\{\vartheta_+(d), \vartheta_{++}(d)\}$$

where  $\vartheta_-(d)$ ,  $\vartheta_+(d)$  and  $\vartheta_{++}(d)$  are given by

$$\begin{aligned} \vartheta_-(d) &= \sqrt[4]{\frac{d^{2d}}{(d+1)^{d+1}(d-1)^{d-1}}} \vartheta_{++}(d), \\ \frac{1}{\vartheta_+(d)} &= \frac{d-1}{d} I_{\frac{d+1}{2d}}\left(\frac{d+1}{4}, \frac{d+3}{4}\right) + \frac{d+1}{d} I_{\frac{d-1}{2d}}\left(\frac{d-1}{4}, \frac{d+5}{4}\right) - 1 \text{ and} \\ \vartheta_{++}(d) &= \sqrt{\frac{\pi}{2}} \frac{\Gamma(\frac{d+3}{2})}{\Gamma(\frac{d}{2} + 1)}. \end{aligned}$$

**13.1. Proof of Theorem 13.1.** Our starting point is the optimization Problem 9.1 (or its equivalent Problem 12.1) from Section 9. Any good approximation  $p$  of  $\sigma_{s,t}$  will now give upper bound  $f_{s,t}(p)$  on the minimum  $f_{s,t}(\sigma_{s,t})$  of  $f_{s,t}$ . This will be our strategy to get a good upper bound of  $\frac{1}{\vartheta(d)}$ , i.e., a good lower bound of  $\vartheta(d)$ . Getting good upper bounds of  $\vartheta(d)$  will be harder. To do this, we introduce sort of artificially simplified versions  $g_{s,t}, h_{s,t}: [0, 1] \rightarrow \mathbb{R}$  of  $f_{s,t}$  having the property

$$f(\sigma_{s,t}) = g(\sigma_{s,t}) = h(\sigma_{s,t})$$

but decreasing at least in one direction while moving away from  $\sigma_{s,t}$ . The upper bound of  $\vartheta(d)$  arising from  $g$  seems to be tighter while the one arising from  $h$  will be given by a simpler expression. Let these functions be given by

$$\begin{aligned} g_{s,t}(p) &= 2 \left( \frac{s I_{1-p}\left(\frac{t}{2}, \frac{s}{2} + 1\right)}{s+t} + \frac{t I_p\left(\frac{s}{2}, \frac{t}{2} + 1\right)}{s+t} \right) - 1 \quad \text{and} \\ h_{s,t}(p) &= I_{1-p}\left(\frac{t}{2}, \frac{s}{2} + 1\right) + I_p\left(\frac{s}{2}, \frac{t}{2} + 1\right) - 1 \end{aligned}$$

for  $p \in [0, 1]$ . Using the standard identities with beta functions, it is easy to compute

$$\begin{aligned} g'_{s,t}(p) &= \frac{2p^{\frac{s}{2}-1}(1-p)^{\frac{t}{2}-1}}{B\left(\frac{s}{2}, \frac{t}{2}\right)}(1-2p) \quad \text{and} \\ h'_{s,t}(p) &= \frac{p^{\frac{s}{2}-1}(1-p)^{\frac{t}{2}-1}(s+t)((1-p)s-pt)}{stB\left(\frac{s}{2}, \frac{t}{2}\right)} \end{aligned}$$

for  $p \in [0, 1]$ . Therefore  $g_{s,t}$  is strictly increasing on  $[0, \frac{1}{2}]$  and strictly decreasing on  $[\frac{1}{2}, 1]$ , and  $h_{s,t}$  is strictly increasing on  $[0, \frac{s}{s+t}]$  and strictly decreasing on  $[\frac{s}{s+t}, 1]$ . Another useful identity which we will use is

$$(13.2) \quad f_{s,t}\left(\frac{s}{s+t}\right) = h_{s,t}\left(\frac{s}{s+t}\right).$$

**Lemma 13.2.** *Let  $s, t \in \mathbb{N}$  and  $p \in [0, 1]$ . Then*

$$\begin{aligned} h_{s,t}(p) &= \frac{2(s+t)}{stB(\frac{s}{2}, \frac{t}{2})} p^{\frac{s}{2}} (1-p)^{\frac{t}{2}} \\ &= \frac{p^{\frac{s}{2}} (1-p)^{\frac{t}{2}}}{(\frac{s}{2} + \frac{t}{2} + 1) B(\frac{s}{2} + 1, \frac{t}{2} + 1)} \\ &= \frac{\Gamma(\frac{s}{2} + \frac{t}{2} + 1)}{\Gamma(\frac{s}{2} + 1) \Gamma(\frac{t}{2} + 1)} p^{\frac{s}{2}} (1-p)^{\frac{t}{2}}. \end{aligned}$$

*Proof.* Using the identities  $I_p(x, y) = I_p(x-1, y+1) - \frac{p^{x-1}(1-p)^y}{yB(x, y)}$ <sup>6</sup> and  $B(\frac{s}{2} + 1, \frac{t}{2}) = \frac{s}{s+t} B(\frac{s}{2}, \frac{t}{2})$ , we get

$$\begin{aligned} 1 - I_{1-p}\left(\frac{t}{2}, \frac{s}{2} + 1\right) &= I_p\left(\frac{s}{2} + 1, \frac{t}{2}\right) = I_p\left(\frac{s}{2}, \frac{t}{2} + 1\right) - \frac{p^{\frac{s}{2}}(1-p)^{\frac{t}{2}}}{\frac{t}{2}B(\frac{s}{2} + 1, \frac{t}{2})} \\ &= I_p\left(\frac{s}{2}, \frac{t}{2} + 1\right) - \frac{2(s+t)p^{\frac{s}{2}}(1-p)^{\frac{t}{2}}}{stB(\frac{s}{2}, \frac{t}{2})} \end{aligned}$$

and therefore  $h_{s,t}(p)$  equals the first of the three expressions. Using  $B(\frac{s}{2} + 1, \frac{t}{2} + 1) = \frac{s}{s+t+2} B(\frac{s}{2}, \frac{t}{2} + 1) = \frac{st}{(s+t)(s+t+2)} B(\frac{s}{2}, \frac{t}{2})$ , we get from this that  $h_{s,t}(p)$  also equals the second expression. Finally,

$$B\left(\frac{s}{2} + 1, \frac{t}{2} + 1\right) = \frac{\Gamma(\frac{s}{2} + 1) \Gamma(\frac{t}{2} + 1)}{\Gamma(\frac{s}{2} + \frac{t}{2} + 2)} = \frac{\Gamma(\frac{s}{2} + 1) \Gamma(\frac{t}{2} + 1)}{(\frac{s}{2} + \frac{t}{2} + 1) \Gamma(\frac{s}{2} + \frac{t}{2} + 1)},$$

yielding that  $h_{s,t}(p)$  equals the third expression. ■

*Proof of Theorem 13.1.* (a) is clear. Part (b) has already been proven in (1.2) but we shortly give again an argument: If  $d$  is even, we know that

$$\frac{1}{\vartheta(d)} = f_{\frac{d}{2}, \frac{d}{2}}(\sigma_{\frac{d}{2}, \frac{d}{2}}) = h_{\frac{d}{2}, \frac{d}{2}}(\sigma_{\frac{d}{2}, \frac{d}{2}})$$

but obviously  $\sigma_{\frac{d}{2}, \frac{d}{2}} = \frac{1}{2}$  and so by Lemma 13.2

$$\frac{1}{\vartheta(d)} = h_{\frac{d}{2}, \frac{d}{2}}\left(\frac{1}{2}\right) = \frac{\Gamma(\frac{d}{2} + 1)}{\Gamma(\frac{d}{4} + 1)^2 2^{\frac{d}{2}}}.$$

By the Lagrange duplication formula<sup>7</sup> we have

$$\Gamma\left(\frac{d}{2} + 1\right) = \Gamma\left(2\left(\frac{d}{4} + \frac{1}{2}\right)\right) = \frac{1}{\sqrt{\pi} 2^{\frac{d}{2}}} \Gamma\left(\frac{d}{4} + \frac{1}{2}\right) \Gamma\left(\frac{d}{4} + 1\right)$$

which proves part (b).

<sup>6</sup>see (8.17.19) in <http://dlmf.nist.gov/8.17>

<sup>7</sup>see 5.5.5 in <http://dlmf.nist.gov/5.5>

It remains to prove (c) and we suppose from now on that  $d \geq 3$  is odd. Then the defining Equation (9.2) for  $p := \sigma_{\frac{d+1}{2}, \frac{d-1}{2}}$  is the one stated in (c) and we know from Theorem 1.2 that

$$\frac{1}{\vartheta(d)} = f_{\frac{d+1}{2}, \frac{d-1}{2}}(p) = g_{\frac{d+1}{2}, \frac{d-1}{2}}(p) = h_{\frac{d+1}{2}, \frac{d-1}{2}}(p).$$

Using Lemma 13.2, we get

$$\frac{1}{\vartheta(d)} = h_{\frac{d+1}{2}, \frac{d-1}{2}}(p) = \frac{\Gamma\left(\frac{d+1}{4} + \frac{d-1}{4} + 1\right)}{\Gamma\left(\frac{d+1}{4} + 1\right) \Gamma\left(\frac{d-1}{4} + 1\right)} p^{\frac{d+1}{4}} (1-p)^{\frac{d-1}{4}},$$

showing the equality we claim for  $\vartheta(d)$ . From Section 11 we know that

$$\frac{1}{2} \leq p \leq \frac{d+1}{2d}.$$

By the monotonicity properties observed earlier, this implies

$$g_{\frac{d+1}{2}, \frac{d-1}{2}}\left(\frac{d+1}{2d}\right) \leq g_{\frac{d+1}{2}, \frac{d-1}{2}}(p) = \frac{1}{\vartheta(d)}$$

and

$$h_{\frac{d+1}{2}, \frac{d-1}{2}}\left(\frac{1}{2}\right) \leq h_{\frac{d+1}{2}, \frac{d-1}{2}}(p) = \frac{1}{\vartheta(d)}.$$

The first inequality shows by a simple calculation that  $\vartheta(d) \leq \vartheta_+(d)$ .

To show that  $\vartheta(d) \leq \vartheta_{++}(d)$ , it is enough to verify that

$$(13.3) \quad h_{\frac{d+1}{2}, \frac{d-1}{2}}\left(\frac{1}{2}\right) = \frac{1}{\vartheta_{++}(d)}.$$

To this end, we use the third expression in Lemma 13.2 to obtain

$$h_{\frac{d+1}{2}, \frac{d-1}{2}}\left(\frac{1}{2}\right) = \frac{\Gamma\left(\frac{d}{2} + 1\right)}{\Gamma\left(\frac{d+1}{4} + 1\right) \Gamma\left(\frac{d-1}{4} + 1\right) 2^{\frac{d}{2}}}.$$

By the Lagrange duplication formula, we have

$$\Gamma\left(\frac{d+3}{2}\right) = \Gamma\left(2\left(\frac{d+3}{4}\right)\right) = \frac{1}{\sqrt{\pi}} 2^{\frac{d+1}{2}} \Gamma\left(\frac{d+3}{4}\right) \Gamma\left(\frac{d+5}{4}\right)$$

and therefore

$$h_{\frac{d+1}{2}, \frac{d-1}{2}}\left(\frac{1}{2}\right) = \frac{\Gamma\left(\frac{d}{2} + 1\right)}{\Gamma\left(\frac{d+3}{2}\right) \frac{\sqrt{\pi}}{\sqrt{2}}} = \sqrt{\frac{2}{\pi}} \frac{\Gamma\left(\frac{d}{2} + 1\right)}{\Gamma\left(\frac{d+3}{2}\right)} = \frac{1}{\vartheta_{++}(d)}.$$

Finally, we show that  $\vartheta_-(d) \leq \vartheta(d)$ . This will follow from

$$\frac{1}{\vartheta(d)} = f_{\frac{d+1}{2}, \frac{d-1}{2}}(p) \leq f_{\frac{d+1}{2}, \frac{d-1}{2}}\left(\frac{d+1}{2d}\right) \stackrel{(13.2)}{=} h_{\frac{d+1}{2}, \frac{d-1}{2}}\left(\frac{d+1}{2d}\right)$$

if we can show that

$$h_{\frac{d+1}{2}, \frac{d-1}{2}}\left(\frac{d+1}{2d}\right) = \frac{1}{\vartheta_-(d)}.$$

But this follows from

$$\begin{aligned} \frac{1}{\vartheta_-(d)} &= \left(\frac{d+1}{d}\right)^{\frac{d+1}{4}} \left(\frac{d-1}{d}\right)^{\frac{d-1}{4}} \frac{1}{\vartheta_{++}(d)} = \\ &= \left(\frac{d+1}{d}\right)^{\frac{d+1}{4}} \left(\frac{d-1}{d}\right)^{\frac{d-1}{4}} h_{\frac{d+1}{2}, \frac{d-1}{2}} \left(\frac{1}{2}\right) = h_{\frac{d+1}{2}, \frac{d-1}{2}} \left(\frac{d+1}{2d}\right) \end{aligned}$$

where the second equality stems from 13.3 and the third from Lemma 13.2.  $\blacksquare$

The following table shows the approximate values of  $\vartheta(d)$  and its lower and upper bounds from Theorem 13.1 for  $t \leq 20$

$d$	$\vartheta_-(d)$	$\vartheta(d)$	$\vartheta_+(d)$	$\vartheta_{++}(d)$
1	—	1	—	—
2	—	1.5708	—	—
3	1.73205	1.73482	1.77064	1.88562
4	—	2	—	—
5	2.15166	2.1527	2.17266	2.26274
6	—	2.35619	—	—
7	2.49496	2.49548	2.50851	2.58599
8	—	2.66667	—	—
9	2.79445	2.79475	2.80409	2.87332
10	—	2.94524	—	—
11	3.064	3.06419	3.07131	3.13453
12	—	3.2	—	—
13	3.31129	3.31142	3.31707	3.37565
14	—	3.43612	—	—
15	3.54114	3.54123	3.54585	3.6007
16	—	3.65714	—	—
17	3.75681	3.75688	3.76076	3.8125
18	—	3.86563	—	—
19	3.96068	3.96073	3.96404	4.01316
20	—	4.06349	—	—

**13.2. Explicit bounds on  $\vartheta(d)$ .** In this subsection we present explicit bounds on  $\vartheta(d)$  for  $d \in \mathbb{N}$ . Theorem 13.1 gives an explicit analytic expression for  $\vartheta(d)$  when  $d \in \mathbb{N}$  is even, and gives analytic bounds for  $\vartheta(d)$  with  $d$  odd.

**Proposition 13.3.** *Let  $d \in \mathbb{N}$ .*

(1) *If  $d$  is even, then*

$$\frac{\sqrt{\pi}}{2} \sqrt{d+1} \leq \vartheta(d) \leq \frac{\sqrt{\pi}}{2} \cdot \frac{d}{\sqrt{d-1}}.$$

(2) If  $d$  is odd, then

$$\sqrt[4]{\left(1 - \frac{1}{d+1}\right)^{d+1} \left(1 + \frac{1}{d-1}\right)^{d-1}} \cdot \frac{\sqrt{\pi}}{2} \sqrt{d + \frac{3}{2}} \leq \vartheta(d) \leq \frac{\sqrt{\pi}}{2} \cdot \frac{d+2}{\sqrt{d + \frac{5}{2}}}.$$

(3) We have  $\lim_{d \rightarrow \infty} \frac{\vartheta(d)}{\sqrt{d}} = \frac{\sqrt{\pi}}{2}$ .

*Proof.* Suppose that  $d$  is even. By Theorem 1.2,

$$(13.4) \quad \vartheta(d) = \sqrt{\pi} \frac{\Gamma\left(\frac{d}{4} + 1\right)}{\Gamma\left(\frac{d}{4} + \frac{1}{2}\right)}.$$

Since  $d$  is even, we may apply Chu's inequality (see Remark 12.10), to the the right-hand side of (13.4) and obtain (1).

Similarly, (2) is obtained by applying Chu's inequality to (13.1). Finally, (3) is an easy consequence of (1) and (2).  $\blacksquare$

Observe that these upper bounds are tighter than the bound  $\vartheta(d) \leq \frac{\pi}{2}\sqrt{d}$  given in [B-TN02].

## 14. DILATIONS AND INCLUSIONS OF BALLS

In this section free relaxations of the problem of including the unit ball of  $\mathbb{R}^g$  into a spectrahedron are considered. Here the focus is the dependence of the inclusion scale as a function of  $g$  (rather than  $d$ ). Among the results we identify the worst case inclusion constant as  $g$ . This inclusion constant can be viewed as a symmetric variable matrix version of the quantitative measure  $\alpha(\mathbb{C}^g)$  of the difference between the maximal and minimal operator space structures associated with the unit ball in  $\mathbb{C}^g$  introduced by Paulsen [Pau02] for which the best results give only upper and lower bounds [Pis03].

**14.1. The general dilation result.** Let  $A \in \mathbb{S}_d^g$  be a given  $g$ -tuple and assume  $\mathcal{D}_{L_A}$  is bounded.

**Proposition 14.1.** *Suppose  $\mathcal{D}_{L_A}$  has the property that if  $C \in \mathcal{D}_{L_A}$  and  $1 \leq j \leq g$ , then*

$$\hat{C}_j = (0, \dots, 0, C_j, 0, \dots, 0) \in \mathcal{D}_{L_A}.$$

*If  $C \in \mathcal{D}_{L_A}(n)$ , then there exists a commuting tuple  $T \in \mathcal{D}_{L_A}(gn)$  such that  $C$  dilates to  $gT$ . The estimate is sharp in that  $g$  is the smallest number such that for every  $g$ -tuple  $A$  (satisfying the assumption above) and  $g$ -tuple  $X \in \mathcal{D}_{L_A}$  there exists a commuting  $g$ -tuple  $T$  of symmetric matrices of size  $gn$  such that  $X$  dilates to  $gT$  and the joint spectrum of  $T$  lies in  $\mathcal{D}_{L_A}(1)$ .*



*Proof.* Given  $C$ , let  $T_j = \bigoplus_{k=1}^g T_{jk}$ , where  $T_{jk} = 0$  (the  $n \times n$  zero matrix) if  $j \neq k$  and  $T_{jj} = C_j$ . It is automatic that the  $T_j$  commute. Further,

$$L_A(T) = \bigoplus_{j=1}^g (I - A_j \otimes C_j) \succeq 0,$$

so that  $T \in \mathcal{D}_{L_A}$ . Finally, let  $V : \mathbb{R}^n \rightarrow \mathbb{R}^n \otimes \mathbb{R}^g$  denote the mapping

$$Vh = \frac{1}{\sqrt{g}} \bigoplus_1^g h.$$

It is routine to verify that  $V^*TV = \frac{1}{g}C$ .

The proof of sharpness is more difficult and is completed in Corollary 14.11. ■

**Remark 14.2.** The hypothesis of Proposition 14.1 applies to the matrix cube. When  $d$  is large and  $g$  is small, the proposition gives a better estimate for the matrix cube relaxation than does Theorem 1.6 of Ben-Tal and Nemirovski. More generally, if  $\mathcal{D}_{L_A}$  is **real Reinhardt**, meaning if  $X = (X_1, \dots, X_g) \in \mathcal{D}_{L_A}$ , then all the tuples  $(\pm X_1, \dots, \pm X_g) \in \mathcal{D}_{L_A}$ , then  $\mathcal{D}_{L_A}$  satisfies the hypothesis of Proposition 14.1. Of course any  $\mathcal{D}_{L_B}$  can be embedded in a  $\mathcal{D}_{L_A}$  which satisfies the hypotheses of Proposition 14.1. □

**14.2. Four types of balls.** Let  $\mathbb{B}_g$  denote the unit ball in  $\mathbb{R}^g$ . Here we consider four matrix convex sets each of which, at level 1, equal  $\mathbb{B}_g$ . Two of these we know to be free spectrahedra. A third is for  $g = 2$ , but likely not for  $g \geq 3$ . The remaining one we will prove is not a free spectrahedron as part of a forthcoming paper.

**14.2.1. The OH ball.** The **OH ball** (for operator Hilbert space [Pis03])  $\mathfrak{B}_g^{\text{oh}}$  is the set of tuples  $X = (X_1, \dots, X_g)$  of symmetric matrices such that

$$\sum_{j=1}^g X_j^2 \preceq I.$$

Equivalently, the row matrix  $\begin{pmatrix} X_1 & X_2 & \dots & X_g \end{pmatrix}$  (or its transpose) has norm at most 1. The ball  $\mathfrak{B}_g^{\text{oh}}$  is symmetric about the origin and also satisfies the conditions of Proposition 14.1.

**Example 14.3.** For two variables,  $g = 2$ , the commutability index of  $\mathfrak{B}_2^{\text{oh}}$  is at least  $\frac{1}{\sqrt{2}}$ .

Let

$$C_1 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \quad \text{and} \quad C_2 = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

Evidently  $C = (C_1, C_2) \in \mathfrak{B}_2^{\text{oh}}$ . Suppose  $T = (T_1, T_2)$  is a commuting tuple of size  $2 + k$  which dilates  $C$ . Thus,

$$T_j = \begin{pmatrix} C_j & a_j \\ a_j^* & d_j \end{pmatrix},$$

where  $a_j$  is  $2 \times k$  and  $d_j$  is  $k \times k$ . Commutativity of the tuple  $T$  implies,

$$\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} = C_1 C_2 - C_2 C_1 = a_1 a_2^* - a_2 a_1^* = \begin{pmatrix} a_1 & -a_2 \end{pmatrix} \begin{pmatrix} a_1^* \\ a_2^* \end{pmatrix}.$$

It follows that either the norm of  $\begin{pmatrix} a_1 & -a_2 \end{pmatrix}$  or  $\begin{pmatrix} a_1 & a_2 \end{pmatrix}^*$  is at least one. In either case,

$$a_1 a_1^* + a_2 a_2^* \not\leq I.$$

On the other hand, the  $(1, 1)$  block entry of  $T_1^2 + T_2^2$  is

$$I_2 + a_1 a_1^* + a_2 a_2^* \not\leq 2I_2.$$

Thus, for the tuple  $C$  the smallest  $\rho$  for which there exists a tuple  $T$  of commuting operators with spectrum in  $\mathbb{B}_g$  such that  $C$  dilates to  $\rho T$  is at least  $\sqrt{2}$ . In other words, the commutability index of  $\mathfrak{B}_2^{\text{oh}}(2)$  is at most  $\frac{1}{\sqrt{2}}$ .  $\square$

**14.2.2. The min and max balls.** Let  $\mathfrak{B}_g^{\min}$  denote the **min ball** (the unit ball  $\mathbb{B}_g$  with the minimum operator system structure). Namely,  $X = (X_1, \dots, X_g) \in \mathfrak{B}_g^{\min}$  if

$$\sum_{j=1}^g x_j X_j \preceq I$$

for all unit vectors  $x \in \mathbb{R}^g$ .

**Lemma 14.4.** *For a tuple  $X$  of  $n \times n$  symmetric matrices, the following are equivalent.*

- (i)  $X$  is in the min ball;
- (ii)  $\mathbb{B}_g \subseteq \mathcal{D}_{L_X}(1)$ ;
- (iii) for each unit vector  $v \in \mathbb{R}^n$  the vector  $v^* X v = (v^* X_1 v, \dots, v^* X_g v) \in \mathbb{B}_g$ ;
- (iv)  $X \in \mathcal{D}_{L_A}(n)$  for every  $g$ -tuple  $A$  of symmetric  $1 \times 1$  matrices for which  $\mathcal{D}_{L_A}(1) \supseteq \mathbb{B}_g$ .

*Proof.* The equivalence of (i) and (ii) is immediate. The tuple  $X$  is in the min ball if and only if for each pair of unit vectors  $x, v \in \mathbb{B}_g$ ,

$$\sum_{j=1}^g x_j (v^* X_j v) \leq v^* v = 1$$

and the equivalence of (i) and (iii) follows. Now suppose  $X$  is not in the min ball. In this case there exists  $a \in \mathbb{B}_g$  such that  $X \notin \mathcal{D}_{L_a}(1)$  but of course  $\mathcal{D}_{L_a}(1) \supseteq \mathbb{B}_g$ . Thus (iv) implies (i). If (iv) doesn't hold, then there is a  $a \in \mathbb{B}_g$  such that  $X \notin \mathcal{D}_{L_a}(1)$ , but  $\mathcal{D}_{L_a}(1) \supseteq \mathbb{B}_g$ . This latter inclusion implies  $a \in \mathbb{B}_g$  and it follows that  $X$  is not in the min ball and (i) implies (iv).  $\blacksquare$

**Remark 14.5.** Note that  $\mathfrak{B}_g^{\min}$  is not exactly a free spectrahedron since it is defined by infinitely many linear matrix inequalities (LMIs). In a forthcoming paper we show using the theory of matrix extreme points that in fact  $\mathfrak{B}_g^{\min}$  is not a free spectrahedron.  $\square$

By comparison, the **max ball**, denoted  $\mathfrak{B}_g^{\max}$ , is the set of  $g$ -tuples of symmetric matrices  $X = (X_1, \dots, X_g)$  such that  $X \in \mathcal{D}_{L_A}$  for every  $d$  and  $g$ -tuple  $A$  of symmetric  $d \times d$  matrices for which  $\mathcal{D}_{L_A}(1) \supseteq \mathbb{B}_g$ . Like the min ball, the max ball is not presented as a free spectrahedron

since it is defined in terms of infinitely many LMIs. It is described by an LMI when  $g = 2$ . See Subsection 14.2.3 below.

**Proposition 14.6.** *The min and max balls are polar dual in the following sense. A tuple  $X \in \mathfrak{B}_g^{\max}$  (resp.  $\mathfrak{B}_g^{\min}$ ) if and only if*

$$(14.1) \quad \sum_{j=1}^g X_j \otimes Y_j \preceq I$$

for every  $Y \in \mathfrak{B}_g^{\min}$  (resp.  $\mathfrak{B}_g^{\max}$ ). Moreover, if  $A$  is a  $g$ -tuple of symmetric  $d \times d$  matrices and if  $\mathcal{D}_{L_A}(1) = \mathbb{B}_g$ , then, for each  $n$ ,

$$\mathfrak{B}_g^{\max}(n) \subseteq \mathcal{D}_{L_A}(n) \subseteq \mathfrak{B}_g^{\min}(n).$$

*Proof.* Recall that  $Y \in \mathfrak{B}_g^{\min}$  if and only if  $\mathcal{D}_{L_Y}(1) \supseteq \mathbb{B}_g$ . Thus  $X \in \mathfrak{B}_g^{\max}$  if and only if  $X \in \mathcal{D}_{L_Y}$  for every  $Y \in \mathfrak{B}_g^{\min}$ . Conversely,  $X \in \mathfrak{B}_g^{\max}$  if and only if  $X \in \mathcal{D}_{L_Y}$  for every  $Y$  such that  $\mathbb{B}_g \subseteq \mathcal{D}_{L_Y}(1)$  (equivalently  $Y \in \mathfrak{B}_g^{\min}$ ).

Now suppose  $\mathbb{B}_g = \mathcal{D}_{L_A}(1)$ . By definition, if  $X \in \mathfrak{B}_g^{\max}$ , then  $X \in \mathcal{D}_{L_A}$  since  $\mathbb{B}_g \subseteq \mathcal{D}_{L_A}(1)$ . On the other hand, if  $X \in \mathcal{D}_{L_A}(n)$ , then, for each unit vector  $v \in \mathbb{R}^n$ ,

$$I = v^* v I \succeq (v^* \otimes I) \left( \sum X_j \otimes A_j \right) (v^* \otimes I) = \sum_j (v^* X_j v) A_j.$$

Hence  $v^* X v \in \mathcal{D}_{L_A}(1) \subseteq \mathbb{B}_g$ . By Lemma 14.4(iii),  $X$  is in the min ball. ■

**Remark 14.7.** The use of the term *minimal* in *min ball* refers not to the size of the spectrahedron itself, but rather by analogy to its use in the theory of operator spaces [Pau02]. In particular, the min ball is defined, in a sense, in terms of a minimal number of positivity conditions; whereas the max ball is defined in terms of the maximal number of positivity conditions (essentially that it should be contained in every free spectrahedron which, at level 1, is the unit ball). Proposition 14.6 is a version of the duality between the minimum and maximum operator space structures on a normed vector space [Pau02]. □

14.2.3. *The spin ball and the canonical anticommutation relations.* Fix a positive integer  $g$ . The description of our fourth ball uses the **canonical anticommutation relations (CAR)** [GåWg54, Der06]. A  $g$ -tuple  $p = (p_1, \dots, p_g)$  of symmetric matrices satisfies the CARs or is a **spin system** [Pis03] if

$$p_j p_k + p_k p_j = 2\delta_{jk} I.$$

One construction of such a system  $P = (P_1, \dots, P_g)$ , and the one adopted here, starts with the spin matrices,

$$\sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \quad \sigma_2 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

For convenience let  $\sigma_0 = I_2$ . Given  $g \geq 2$ , each  $P_j$  is a  $(g-1)$ -fold tensor product of combinations of the  $2 \times 2$  matrices  $\sigma_j$  for  $0 \leq j \leq 2$ . In particular, each  $P_j$  is a symmetric matrix of size  $2^{g-1}$ . Define  $P_1 = \sigma_1 \otimes \sigma_0 \otimes \dots \otimes \sigma_0$  and, for  $2 \leq j \leq g-1$ ,

$$P_j = \sigma_2 \otimes \dots \otimes \sigma_2 \otimes \sigma_1 \otimes \sigma_0 \otimes \dots \otimes \sigma_0,$$

where  $\sigma_1$  appears in the  $j$ -th (reading from the left) tensor; thus  $\sigma_2$  appears  $j - 1$  times and  $\sigma_0$  appears  $g - j - 1$  times. Finally, let  $P_g$  denote the  $(g - 1)$ -fold tensor product  $\sigma_2 \otimes \cdots \otimes \sigma_2$ .

The **spin ball**, denoted  $\mathfrak{B}_g^{\text{spin}}$ , is the free spectrahedron determined by the tuple  $P$ . Thus,  $X \in \mathfrak{B}_g^{\text{spin}}$  if and only if  $L_P(X) = I - \sum P_j \otimes X_j \succeq 0$  if and only if  $\sum_{j=1}^g X_j \otimes P_j$  is a contraction (cf. Lemma 14.8(iii)). Further relations between the three balls are explored in the following subsection.

Here are some further observations on the spin ball. Let

$$\sigma_3 = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$$

and let

$$\Sigma = \Sigma_g = \left\{ \bigotimes_{j=1}^{g-1} \sigma_{j_k} : j_k \in \{0, 1, 2, 3\} \right\}.$$

In particular, the cardinality of  $\Sigma_g$  is  $4^{g-1}$ .

**Lemma 14.8.** *If  $X$  is a  $g$ -tuple of  $d \times d$  symmetric matrices, then the matrix  $\sum_{j=1}^g X_j \otimes P_j$  has the following properties.*

(i)

$$(I \otimes u)^* \left( \sum_{j=1}^g X_j \otimes P_j \right) (I \otimes u) = \sum_{j=1}^g \pm X_j \otimes P_j$$

for each  $u \in \Sigma$  with each sequence of  $\pm$  assumed  $2^{g-2}$  times;

(ii) the sets  $\Sigma_g^\pm$  consisting of those elements  $u$  of  $\Sigma_g$  such that  $(I \otimes u)^* (\sum_{j=1}^g X_j \otimes P_j) (I \otimes u) = \pm \sum_{j=1}^g X_j \otimes P_j$  have  $2^{g-2}$  elements; each  $u \in \Sigma_g^\pm$ , other than the identity, is skew symmetric;

(iii)  $\sum_j X_j \otimes P_j$  is unitarily equivalent to  $-\sum_j X_j \otimes P_j$ ;

(iv)  $\sum_j X_j \otimes P_j$  has  $2d$  eigenvalues (coming in  $d$  pairs of  $\pm \lambda$  by item (iii)) each with multiplicity  $2^{g-2}$ ;

(v) If  $(\sum_j X_j \otimes P_j) \Gamma = 0$ , then  $(\sum_j X_j \otimes P_j) (I \otimes u) \Gamma = 0$  for all  $u \in \Sigma_g^+ \cup \Sigma_g^-$ .

*Proof.* We first prove

$$(14.2) \quad \sigma_k \sigma_j \sigma_k = \pm \sigma_j$$

for  $0 \leq j, k \leq 3$ . Observe it may be assumed that  $1 \leq j, k \leq 3$ . For such  $j, k$ , with  $s \in \{1, 2, 3\} \setminus \{j, k\}$ , evidently  $\sigma_j \sigma_k = \sigma_s$ . Hence,  $\sigma_k \sigma_s = \sigma_j$  (since  $j \in \{1, 2, 3\} \setminus \{k, s\}$ ) and Equation (14.2) follows.

Equation (14.2) immediately implies  $u^* P_k u = \pm P_k$  for  $u \in \Sigma_g$  and  $1 \leq k \leq g$ . Thus  $(I \otimes u)^* [\sum_j X_j \otimes P_j] (I \otimes u) = \sum_{j=1}^g \pm X_j \otimes P_j$  for some choice of signs. This proves the first part of item (i). The rest of item (i) is established after the proof of item (ii).

Turning to the proof of item (ii), observe  $u \in \Sigma_g^+$  if and only if  $u^* P_k u = P_k$  for each  $1 \leq k \leq g$ . When  $g = 2$ ,  $P_1 = \sigma_1$  and  $P_2 = \sigma_2$ . Evidently,  $\Sigma_2^+ = \{\sigma_0\}$  and  $\Sigma_2^- = \{\sigma_3\}$ . Now suppose item (ii) holds for  $g$ . In this case, letting  $\{P_1, \dots, P_g\}$  denote the CAR matrices for  $g$ , the CAR matrices for  $g + 1$  are  $\{q_1, \dots, q_{g+1}\}$  where  $q_1 = \sigma_1 \otimes 1$  and  $q_j = \sigma_2 \otimes P_{j-1}$  for  $j > 1$ .

If  $u$  is in  $\Sigma_g^+$ , then  $I \otimes u \in \Sigma_{g+1}^+$  and  $\sigma_3 \otimes u \in \Sigma_{g+1}^-$ . Similarly, if  $u \in \Sigma_g^-$ , then  $\sigma_1 \otimes u \in \Sigma_{g+1}^+$  and  $\sigma_2 \otimes u \in \Sigma_{g+1}^-$ . It follows that  $\Sigma_{g+1}^-$  has at least  $2^{g-1}$  elements and all of these, except for the identity are skew symmetric. Since  $v = \sigma_3 \otimes I \in \Sigma_{g+1}^-$ , it follows that  $vu \in \Sigma_{g+1}^-$  for each  $u \in \Sigma_{g+1}^+$  and therefore  $\Sigma_{g+1}^-$  has at least  $2^{g-1}$  elements too. By induction, item (ii) holds. Moreover, this argument shows there is a positive integer  $N_g$  so that each sign arrangement is taken either 0 or  $N_g$  times.

We now use induction to show that every sign arrangement is assumed and thus complete the proof of item (i). The result is evident for  $g = 2$ . Now assume it is true for  $g$ . Let  $r = \sigma_3 \otimes \sigma_3 \otimes I$ . Compute  $r^*q_1r = \sigma_3^*\sigma_1\sigma_3 \otimes (\sigma_3^*I\sigma_3) = -q_1$ , but

$$\begin{aligned} r'q_{j+1}r &= \sigma_3^*\sigma_2\sigma_3 \otimes (\sigma_3 \otimes I)P_j(\sigma_3 \otimes I) \\ &= -\sigma_2 \otimes (-P_j) = q_{j+1}. \end{aligned}$$

Thus the combinations of  $I \otimes u$  and  $r^*(I \otimes u)$  for  $u \in \Sigma_g$  produce all sign combinations for  $g + 1$ .

Item (iii) is an immediate consequence of item (ii).

To prove item (iv) - namely to see that each eigenvalue has multiplicity a multiple of  $2^{g-2}$  - observe if  $(\sum X_j \otimes P_j)\Gamma = \lambda\Gamma$ , then the set  $\{u\Gamma : u \in \Sigma_g^+\}$  is linearly independent. To verify this last assertion, first note that if  $u, v \in \Sigma_g^+$ , then  $uv \in \pm\Sigma_g^+$ . Further, each  $u \in \Sigma_g^+$  is skew symmetric, except for the identity 1. In particular,

$$\langle (1 \otimes u)\Gamma, \Gamma \rangle = 0$$

for  $u \neq 1$ . Hence, if  $\sum_{u \in \Sigma_g^+} c_u(I \otimes u)\Gamma = 0$ , then by multiplying by a  $v$  for which  $c_v \neq 0$ , we can assume  $c_1$  (the constant corresponding to the identity 1) is not zero. Thus,

$$0 = \langle \sum c_u(1 \otimes u)\Gamma, \Gamma \rangle = c_1 \|\Gamma\|^2$$

and a contradiction is obtained. To complete the proof, let  $m$  is the largest integer such that there exists  $\Gamma^1, \dots, \Gamma^m$  such that  $\{u\Gamma^k : u \in \Sigma_g^+, 1 \leq k \leq m\}$  spans the eigenspace corresponding to eigenvalue  $\lambda$ , then the dimension of this space is  $m2^{g-2}$ . We prove this assertion using induction. Suppose  $1 \leq k \leq m$  and  $S_k := \{u\Gamma^j : u \in \Sigma_g^+, 1 \leq j \leq k\}$  is linearly independent (and thus has dimension  $k2^{g-2}$ ). Arguing by contradiction, suppose the  $\Delta$  is in the intersection of  $S_k$  and  $\{u\Gamma^{k+1} : u \in \Sigma_g^+\}$ . From what is already been proved, the dimension of the span of  $\{u\Delta : u \in \Sigma_g^+\}$  is  $2^{g-2}$  but this subspace is a subspace of both the span of  $S_k$  and the span of  $\{u\Gamma^{k+1} : u \in \Sigma_g^+\}$ , contradicting the minimality of  $m$ . Hence the dimension of the span of  $S_{k+1}$  is  $(k+1)2^{g-2}$  and the proof is complete.

Item (v) is evident. ■

**Lemma 14.9.** *The mapping from  $\mathbb{R}^g$  (in the Euclidean norm) to  $M_{2^{g-1}}(\mathbb{R})$  (in the operator norm),*

$$x \mapsto \sum_{j=1}^g x_j P_j$$

*is an isometry. In particular,  $\mathfrak{B}_g^{\text{spin}}(1) = \mathbb{B}_g$ .*

*Proof.* The result follows from  $(\sum_{j=1}^g x_j P_j)^2 = (\sum_j x_j^2)I + \sum_{j < k} (x_j x_k - x_k x_j) P_j P_k = (\sum x_j^2)I$ .  $\blacksquare$

A consequence of Lemma 14.9 is that the tuple  $(P_1, \dots, P_g)$  is in  $\mathfrak{B}_g^{\min}$ .

**14.3. Inclusions and dilations.** In this subsection we investigate inclusions between the different types of balls introduced above.

**Lemma 14.10.** *The norm of  $\sum_{j=1}^g P_j \otimes P_j$  is  $g$ . Hence  $P$  is in the (topological) boundary of  $g\mathfrak{B}_g^{\text{spin}}$ . The norm of the block row matrix  $\begin{pmatrix} P_1 & \cdots & P_g \end{pmatrix}$  is  $\sqrt{g}$ . Thus  $P$  is in the boundary of  $\sqrt{g}\mathfrak{B}_g^{\text{oh}}$ .*

*Proof.* We prove a bit more. Let  $\{e_0, e_1\}$  denote the standard basis of  $\mathbb{R}^2$ . In particular,  $\sigma_0 e_j = e_j$ ,  $\sigma_1 e_j = (-1)^j e_j$  and  $\sigma_2 e_j = e_{j+1}$  (modulo 2). For convenience, let  $h = g - 1$ . Given  $\alpha = (\alpha_1, \dots, \alpha_h) \in \mathbb{Z}_2^h$ , let

$$e_\alpha = e_{\alpha_1} \otimes \cdots \otimes e_{\alpha_h} \in \mathbb{R}^{2^h}$$

and

$$\gamma = \sum_{\alpha \in \mathbb{Z}_2^h} e_\alpha \otimes e_\alpha.$$

Now we verify that, for  $1 \leq j \leq g$ ,

$$(14.3) \quad P_j \otimes P_j \gamma = \gamma.$$

Indeed,  $P_1 e_\alpha = (-1)^{\alpha_1} e_\alpha$  and hence  $P_1 \otimes P_1 e_\alpha \otimes e_\alpha = e_\alpha \otimes e_\alpha$ . For  $2 \leq j \leq h$ ,

$$\begin{aligned} P_j e_\alpha &= e_{\alpha_1} \otimes \cdots \otimes e_{\alpha_h} \\ &= e_{\alpha_1+1} \otimes \cdots \otimes e_{\alpha_{j-1}+1} \otimes (-1)^{\alpha_j} e_{\alpha_j} \otimes e_{\alpha_{j+1}} \otimes \cdots \otimes e_{\alpha_h} \\ &= (-1)^{\alpha_j} e_\beta, \end{aligned}$$

where  $\beta = (\alpha_1 + 1, \dots, \alpha_{j-1} + 1, \alpha_j, \dots, \alpha_h)$ . Thus,

$$(P_j \otimes P_j)(e_\alpha \otimes e_\alpha) = e_\beta \otimes e_\beta$$

and the conclusion  $P_j \otimes P_j \gamma = \gamma$  for  $1 \leq j \leq h$  follows. The argument that  $P_g \otimes P_g \gamma = \gamma$  is similar.

It follows that  $(\sum_j P_j \otimes P_j) \gamma = g \gamma$  and hence the norm of  $\sum P_j \otimes P_j$  is at least  $g$ . Since the norm of each  $P_j$  is one, the norm of  $\sum P_j \otimes P_j$  is at most  $g$ . The remainder of the lemma is evident.  $\blacksquare$

**Corollary 14.11.** *The smallest  $\rho$  such that  $\mathfrak{B}_g^{\min} \subseteq \rho \mathfrak{B}_g^{\text{spin}}$  is  $\rho = g$ . In particular, the estimate  $g$  in Proposition 14.1 is sharp.*

*Proof.* Suppose  $\mathfrak{B}_g^{\min} \subseteq \rho \mathfrak{B}_g^{\text{spin}}$ . By Lemma 14.9, the tuple  $P$  coming from the CAR is in  $\mathfrak{B}_g^{\min}$  and by Lemma 14.10,  $P$  is in the boundary of  $g\mathfrak{B}_g^{\text{spin}}$ . Thus  $\rho \geq g$ . On the other hand, if  $X \in \mathfrak{B}_g^{\min}$ , then, since  $\mathfrak{B}_g^{\min}$  satisfies the hypotheses of Proposition 14.1,  $X = gV^*TV$ , where  $V$  is an isometry and  $T$  is a commuting tuple of self adjoint matrices with spectrum in  $\mathbb{B}_g$ . In particular,  $T \in \mathfrak{B}_g^{\text{spin}}$  and thus  $\frac{1}{g}X \in \mathfrak{B}_g^{\text{spin}}$  too. Hence  $\mathfrak{B}_g^{\min} \subseteq g\mathfrak{B}_g^{\text{spin}}$ . Further, if the

estimate in Proposition 14.1 were not sharp, the argument just given would produce a  $\rho < g$  such that  $\mathfrak{B}_g^{\min} \subseteq \rho \mathfrak{B}_g^{\text{spin}}$ , a contradiction. ■

**Theorem 14.12.** The smallest  $\rho$  such that  $\mathfrak{B}_g^{\text{oh}}$  embeds into  $\rho \mathfrak{B}_g^{\text{spin}}$  is  $\sqrt{g}$ .

*Proof.* By Lemma 14.10,  $P$  is in the topological boundaries of  $\sqrt{g} \mathfrak{B}_g^{\text{oh}}$  and  $g \mathfrak{B}_g^{\text{spin}}$ . Hence  $\rho \geq \sqrt{g}$ .

To prove the converse inequality, we use complete positivity to show  $\mathfrak{B}_g^{\text{oh}} \subseteq \sqrt{g} \mathfrak{B}_g^{\text{spin}}$ . We follow the solution [HKM13] to the free spectrahedral inclusion problem as described in Subsection 1.3.2. Let  $e_{i,j}$  denote the  $(g+1) \times (g+1)$  matrix units. Letting  $A_i = e_{1,i+1} + e_{i+1,1}$  for  $i = 1, \dots, g$ , and  $A = (A_1, \dots, A_g)$ , we have  $\mathfrak{B}_g^{\text{oh}} = \mathcal{D}_A$ . Similarly,  $\mathfrak{B}_g^{\text{spin}} = \mathcal{D}_P$ , where  $P = (P_1, \dots, P_g)$ . It thus suffices to show there is a unital completely positive map

$$\psi : e_{1,i+1} + e_{i+1,1} \mapsto \frac{1}{\sqrt{g}} P_i, \quad i = 1, \dots, g.$$

Consider the following ansatz for the Choi matrix for  $\psi$ :

$$(14.4) \quad C_\psi = \begin{pmatrix} \frac{1}{2}I & \frac{1}{2\sqrt{g}}P_1 & \cdots & \frac{1}{2\sqrt{g}}P_g \\ \frac{1}{2\sqrt{g}}P_1 & & & \\ \vdots & & S & \\ \frac{1}{2\sqrt{g}}P_g & & & \end{pmatrix}$$

Set

$$S = \frac{1}{2g} \begin{pmatrix} P_1 \\ \vdots \\ P_g \end{pmatrix} \begin{pmatrix} P_1 \\ \vdots \\ P_g \end{pmatrix}^* = \frac{1}{2g} \begin{pmatrix} I & P_1 P_2 & \cdots & P_1 P_g \\ P_2 P_1 & I & \cdots & P_2 P_g \\ \vdots & \ddots & \ddots & \vdots \\ P_g P_1 & \cdots & \cdots & I \end{pmatrix}$$

It is clear that  $S \succeq 0$ . Furthermore, the Schur complement of the top left block of  $C_\psi$  from (14.4) is 0. Thus  $C_\psi$  is positive semidefinite. Furthermore,  $\frac{1}{2}I + g\frac{1}{2g}I = I$ , whence  $\psi$  is unital. ■

**Proposition 14.13.** The smallest  $\rho$  such that  $\mathfrak{B}_g^{\min}$  embeds into  $\rho \mathfrak{B}_g^{\text{oh}}$  is  $\rho = \sqrt{g}$ .

*Proof.* The tuple  $P$  is in  $\mathfrak{B}_g^{\min}$ , but is, by Lemma 14.10, in the topological boundary of  $\sqrt{g} \mathfrak{B}_g^{\text{oh}}$ . Thus  $\rho \geq \sqrt{g}$ . On the other hand, if  $X = (X_1, \dots, X_g)$  is in  $\mathfrak{B}_g^{\min}$ , then each  $X_j$  is a contraction. Hence the norm of the row matrix  $X = (X_1 \ \cdots \ X_g)$  is at most  $\sqrt{g}$ ; i.e.,  $X \in \mathfrak{B}_g^{\text{oh}}$ . ■

**Proposition 14.14.** A 2-tuple  $X$  is in the spin ball  $\mathfrak{B}_2^{\text{spin}}$  if and only if it dilates to a commuting 2-tuple  $T$  of symmetric matrices (an upper bound on the size of the matrices in  $T$  in terms of  $g$  and  $d$  can be deduced from the proof) with joint spectrum in  $\mathbb{B}_2$ .

Before giving the proof of Proposition 14.14 let us note a few consequences.

**Corollary 14.15.**  $\mathfrak{B}_2^{\text{spin}} = \mathfrak{B}_2^{\max}$ . In particular, for a given 2-tuple  $A$  of  $d \times d$  matrices,  $\mathbb{B}_2 \subseteq \mathcal{D}_{L_A}(1)$  if and only if  $\mathfrak{B}_2^{\text{spin}} \subseteq \mathcal{D}_{L_A}$ . Finally,  $Y \in \mathfrak{B}_2^{\min}$  if and only if there exists a positive integer  $\mu$  such that  $Y$  dilates to  $I_\mu \otimes P$ .

*Proof.* By Proposition 14.6,  $\mathfrak{B}_2^{\max} \subseteq \mathfrak{B}_2^{\text{spin}}$ . Thus it remains to show if  $X \in \mathfrak{B}_2^{\text{spin}}$ , then  $X \in \mathfrak{B}_2^{\max}$ . By Proposition 14.14, there exists a commuting pair  $T$  of symmetric matrices and an isometry  $V$  such that  $X_j = V^* T_j V$  and the joint spectrum of  $T$  is in  $\mathbb{B}_2$ . It follows that  $T \in \mathfrak{B}_2^{\max}$  and thus  $X \in \mathfrak{B}_2^{\max}$  too.

To prove second statement, note that, by definition of the max ball, if  $\mathbb{B}_2 \subseteq \mathcal{D}_{L_A}(1)$ , then  $\mathfrak{B}_2^{\max} \subseteq \mathcal{D}_{L_A}$ . The converse is automatic since  $\mathfrak{B}_2^{\text{spin}}(1) = \mathbb{B}_2$ . Thus the second part of the corollary follows immediately from the first.

The final part of the corollary follows easily from [HKM+]. Here is a sketch. Let  $\mathcal{P}$  denote the span of  $\{I, P_1, \dots, P_k\}$ . Suppose  $Y \in \mathfrak{B}_2^{\min}$  and let  $\mathcal{Y}$  denote the span of  $\{I, Y_1, Y_2\}$ . By Proposition 14.6 and what has already been proved, if  $X \in \mathfrak{B}_2^{\text{spin}}$ , then  $I \succeq \sum_{j=1}^2 X_j \otimes Y_j$ . Hence the unital mapping  $\varphi : \mathcal{P} \rightarrow \mathcal{Y}$  defined by  $\varphi(P_j) = Y_j$  is completely positive. The dilation conclusion follows. Conversely, if  $Y$  dilates to  $I_\mu \otimes P$ , then evidently  $Y$  is in  $\mathfrak{B}_2^{\min}$  and the proof is complete. ■

**Remark 14.16** (Matrix Ball Problem). Given a  $d \times d$  monic linear pencil  $L$  consider the problem of embedding the unit ball  $\mathbb{B}_g$  into the spectrahedron  $\mathcal{S}_L = \mathcal{D}_L(1)$ . Equivalently (Corollary 14.15), consider embedding  $\mathfrak{B}_g^{\text{spin}}$  into  $\mathcal{D}_L$ . Both objects are free spectrahedra, so the complete positivity machinery on embeddings of free spectrahedra applies (see Subsubsection 1.3.2 or [HKM13] for details). That is,  $\mathbb{B}_g \subseteq \mathcal{S}_L = \mathcal{D}_L(1)$  is equivalent to an explicit LMI of size  $2^{g-1}gd$ . □

The proof of Proposition 14.14 uses the following proposition.

**Proposition 14.17.** *A tuple  $X \in \mathfrak{B}_2^{\text{spin}}(d)$  is an extreme point of  $\mathfrak{B}_2^{\text{spin}}(d)$  if and only if it is a commuting tuple and  $\sum_{j=1}^g X_j \otimes P_j$  is unitary.*

*Proof.* For notational ease, let  $\Lambda(x) = \sum_{j=1}^g P_j x_j$ , with the dependence on  $g$  suppressed. Observe that, by Lemma 14.8, a tuple  $X \in \mathfrak{B}_g^{\text{spin}}$  if and only if  $\Lambda(X)^2 \preceq I$ . Equivalently,  $X$  is in the spin ball if and only if  $L(X) := I - \Lambda_*(X) \succeq 0$ , where

$$\Lambda_*(X) = \Lambda(X) \oplus -\Lambda(X) = \begin{pmatrix} \Lambda(X) & 0 \\ 0 & -\Lambda(X) \end{pmatrix}.$$

In the case of  $g = 2$  and the tuple  $X = (X_1, X_2)$ ,

$$\Lambda(X) = \begin{pmatrix} X_1 & X_2 \\ X_2 & -X_1 \end{pmatrix}$$

and  $X$  is in the spin ball if and only if

$$I \succeq \Lambda(X)^2 = \begin{pmatrix} X_1^2 + X_2^2 & X_1 X_2 - X_2 X_1 \\ -(X_1 X_2 - X_2 X_1) & X_1^2 + X_2^2 \end{pmatrix}.$$

Suppose this inequality holds. If  $\gamma \in \mathbb{R}^d$  and  $(X_1^2 + X_2^2)\gamma = \gamma$ , then evidently  $(X_1 X_2 - X_2 X_1)\gamma = 0$ . Let

$$\mathcal{H} = \{\gamma \in \mathbb{R}^d : (X_1^2 + X_2^2)\gamma = \gamma\}.$$



If  $\mathcal{H} = \mathbb{R}^d$ , then  $X_1$  and  $X_2$  commute. Hence, we may assume  $\mathcal{H}$  is a proper subspace of  $\mathbb{R}^d$ . Let  $\mathcal{K} = \mathbb{R}^d \ominus \mathcal{H}$ .

Let  $\mathcal{E}_\pm$  denote the nullspace of  $I \mp \Lambda(X)$ . In particular,  $\mathcal{E}$ , the nullspace of  $I - \Lambda(X)^2$  is the direct sum  $\mathcal{E}_+ \oplus \mathcal{E}_-$ . Let  $\{e_0, e_1\}$  denote the standard basis for  $\mathbb{R}^2$ . If  $\gamma_0 \in \mathcal{K}$  and  $\gamma_1 \in \mathbb{R}^d$  and  $\Gamma = \sum_{j=0}^1 \gamma_j \otimes e_j \in \mathcal{E}_\pm$ , then, since  $(1 \otimes \sigma_3)\Lambda(x)(1 \otimes \sigma_3) = -\Lambda(x)$  (see Lemma 14.8),

$$\Lambda(X)(I \otimes \sigma_3)\Gamma = \mp(I \otimes \sigma_3)\Gamma,$$

and hence  $(I \otimes \sigma_3)\Gamma$  lies in  $\mathcal{E}_\mp$ . It follows that  $I \otimes \sigma_3$  leaves  $\mathcal{E}$  invariant. Let  $Q$  denote the projection of  $\mathbb{R}^d$  onto  $\mathcal{K}$ . If  $\Gamma \in \mathcal{E}$ , then  $((I - Q) \otimes I)\Gamma \in \mathcal{H} \otimes \mathbb{R}^2 \subseteq \mathcal{E}$ . Hence  $(Q \otimes I)\Gamma \in \mathcal{E}$ . Consequently, we can assume there is a nonzero  $\Gamma \in \mathcal{E} \cap (\mathcal{K} \otimes \mathcal{K})$ . Write, as before  $\Gamma = \sum_{j=0}^1 \gamma_j \otimes e_j$ . To see that  $\{\gamma_0, \gamma_1\}$  spans a two dimensional space  $\mathcal{S}$  (is a linearly independent set), suppose not. In that case,  $\Gamma = \gamma \otimes e$  for vectors  $\gamma \in \mathcal{K}$  and  $e \in \mathbb{R}^2$ . Since  $\sigma_3$  is skew symmetric (as is  $X_1X_2 - X_2X_1$ ),

$$\left\langle [(X_1X_2 - X_2X_1) \otimes \sigma_3]\Gamma, \Gamma \right\rangle = \langle (X_1X_2 - X_2X_1)\gamma, \gamma \rangle \langle \sigma_3 e, e \rangle = 0.$$

Thus, as  $\Gamma = \Lambda(X)^2\Gamma$ ,

$$\begin{aligned} \|\gamma\|^2 \|e\|^2 &= \|\Gamma\|^2 \\ &= \langle \Gamma, \Gamma \rangle \\ &= \langle \Lambda(X)^2\Gamma, \Gamma \rangle \\ &= \langle (X_1^2 + X_2^2)\gamma \otimes e + (X_1X_2 - X_2X_1)\gamma \otimes e, \gamma \otimes e \rangle \\ &= \langle (X_1^2 + X_2^2)\gamma, \gamma \rangle \|e\|^2. \end{aligned}$$

Since  $I \succeq X_1^2 + X_2^2$  it follows that  $(X_1^2 + X_2^2)\gamma = \gamma \in H$ . Hence  $\gamma = 0$ .

Both  $\Gamma$  and  $(I \otimes \sigma_3)\Gamma \in \mathcal{E} \cap (\mathcal{S} \oplus \mathcal{S})$ . On the other hand, if  $\{\Gamma, (I \otimes \sigma_3)\Gamma\}$  don't span  $\mathcal{E} \cap (\mathcal{S} \oplus \mathcal{S})$ , then it is easily shown that this intersection contains an element of the form  $\gamma \otimes e$  and we obtain a contradiction as before. Thus  $\mathcal{E} \cap (\mathcal{S} \oplus \mathcal{S})$  is spanned by  $\{\Gamma, (I \otimes \sigma_3)\Gamma\}$ .

Define  $Y = (Y_1, Y_2)$  on  $\mathcal{S}$  as follows. Let  $Z$  be a generic  $4 \times 4$  matrix. We will choose  $Z$  so that it has the form  $\Lambda(Y)$  and such that  $\Lambda(Z)$  is zero on  $\mathcal{E} \cap (\mathcal{S} \oplus \mathcal{S})$ . There are 16 free variables in  $Z$ . Insuring  $Z$  has the proper form with  $Y_j = Y_j^*$  consists of  $2 + 2 \cdot 4 = 10$  homogeneous equations. The condition  $Z\Gamma = 0$  is another 4 homogeneous equations. The form of  $Z$  and  $Z\Gamma = 0$  implies  $Z(I \otimes \sigma_3)\Gamma = 0$  too. Hence we are left with 2 free variables and a choice of  $Y \neq 0$  exists. And  $\Lambda(Y)\mathcal{E} \cap (\mathcal{S} \oplus \mathcal{S}) = 0$  for such a choice of  $Y$ .

Extend the definition of  $Y$  to all of  $\mathbb{R}^d$  by declaring  $Y_j = 0$  on  $\mathcal{S}^\perp$ . It follows that  $Y \neq 0$  and at the same time  $\Lambda(Y)$  is 0 on  $\mathcal{E}$  and hence on each of  $\mathcal{E}_\pm$ . With respect to the decomposition of the space that  $\Lambda(X)$  acts upon as  $\mathcal{E}_+ \oplus \mathcal{E}_- \oplus H$ ,

$$\Lambda(X) = \begin{pmatrix} I & 0 & 0 \\ 0 & -I & 0 \\ 0 & 0 & X' \end{pmatrix}, \quad \Lambda(Y) = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & Y' \end{pmatrix},$$

where  $X', Y'$  are self-adjoint and  $X'$  is a strict contraction. It follows, by choosing  $t$  small enough, that  $I \pm \Lambda(X + tY) \succeq 0$  and thus  $X$  is not an extreme point of  $\mathfrak{B}_2^{\text{spin}}$ . ■

*Proof of Proposition 14.14.* If  $X$  in the spin ball  $\mathfrak{B}_2^{\text{spin}}(d)$ , then by Caratheodory's Theorem there exists an  $N$ , extreme points  $X^1, \dots, X^N$  of  $\mathfrak{B}_2^{\text{oh}}(d)$  and scalars  $0 \leq t_1, \dots, t_N$  such that  $\sum t_j = 1$  and  $X = \sum t_j X^j$ . Let  $T = \oplus X^j$  act on  $H = \oplus_1^N \mathbb{R}^d$  and define  $V : \mathbb{R}^d \rightarrow H$  by  $Vh = \oplus \sqrt{t_j} h$ . Thus  $V$  is an isometry and it is evident that  $V^*TV = \sum t_j X^j = X$ . By Proposition 14.17,  $T$  is a commuting tuple of symmetric matrices. ■

14.3.1. *An alternate proof of Proposition 14.14.* An ad hoc proof of Proposition 14.14 is based upon the Halmos dilation of a contraction matrix to a unitary matrix.

**Lemma 14.18.** *Suppose  $u, v, a, b, d$  are  $n \times n$  matrices and let*

$$R = \begin{pmatrix} a & b \\ b^* & d \end{pmatrix}$$

*If  $R$  is positive semidefinite, and*

$$R^2 = Z := \begin{pmatrix} u & v \\ -v & u \end{pmatrix},$$

*then  $a = d$  and  $b^* = -b$ .*

*Proof.* Note that  $U^*R^2U = R^2$ , where  $U$  is the unitary matrix

$$U = \frac{1}{2} \begin{pmatrix} I & I \\ -I & I \end{pmatrix}.$$

Using the functional calculus it follows that  $U^*RU = R$  too. From this relation, direct computation reveals

$$\begin{aligned} a - (b + b^*) + d &= 2a \\ a + (b - b^*) - d &= 2b \end{aligned}$$

from which it follows that  $b + b^* = 0$  and  $a = d$ . ■

*Proof of Proposition 14.14.* Let

$$S = \begin{pmatrix} X & Y \\ Y & -X \end{pmatrix}.$$

Almost by definition,  $X \in \mathfrak{B}_2^{\text{spin}}$  means  $S$  is a contraction. Let  $D = (I - S^2)^{\frac{1}{2}}$ , the defect of  $S$ . By Lemma 14.18,

$$D = \begin{pmatrix} d & e \\ -e & d \end{pmatrix}.$$

In particular,  $e$  is skew symmetric,  $e^* = -e$ . The operator

$$U = \begin{pmatrix} S & D \\ D & -S \end{pmatrix}$$

is unitary and consequently

$$(14.5) \quad \begin{aligned} X^2 + Y^2 + d^2 - e^2 &= I \\ XY - YX + de + ed &= 0 \\ Yd + Xe + eX - dY &= 0 \\ Xd - Ye - dX - eY &= 0. \end{aligned}$$

Let

$$T_1 = \begin{pmatrix} X & e \\ -e & X \end{pmatrix}, \quad T_2 = \begin{pmatrix} Y & d \\ d & -Y \end{pmatrix}.$$

Compute, using (14.5),

$$T_1 T_2 - T_2 T_1 = \begin{pmatrix} XY + ed - YX + de & Xd - eY - Ye - dX \\ -eY + Xd - dX - Ye & -ed - XY - de + YX \end{pmatrix} = 0.$$

Likewise,

$$T_1^2 + T_2^2 = \begin{pmatrix} X^2 - e^2 + Y^2 + d^2 & Xe + eX + Yd - dY \\ -eX - Xe + dY - Yd & X^2 - e^2 + Y^2 + d^2 \end{pmatrix} = I. \quad \blacksquare$$

## 15. PROBABILISTIC THEOREMS AND INTERPRETATIONS CONTINUED

This section follows up on Section 1.8, adding a few more probabilistic facts and summarizing properties involving equipoints. We follow the conventions of Section 1.8. In particular, for  $\mathfrak{s}, \mathfrak{t} \in \mathbb{R}$  with  $\mathfrak{d} = \mathfrak{s} + \mathfrak{t} > 0$  the equipoint  $e_{\mathfrak{s}, \mathfrak{t}}$  is defined by

$$(15.1) \quad P^{b(\mathfrak{s}+1, \mathfrak{t})}(\mathfrak{B} \leq e_{\mathfrak{s}, \mathfrak{t}}) = P^{b(\mathfrak{s}, \mathfrak{t}+1)}(\mathfrak{B} \geq e_{\mathfrak{s}, \mathfrak{t}}).$$

**15.1. The nature of equipoints.** Here are basic properties of equipoints versus medians.

**Proposition 15.1.** *Various properties of the distributions  $\text{Bin}(\mathfrak{d}, p)$  and  $\text{Beta}(\mathfrak{s}, \mathfrak{t})$  are:*

- (1) *Bin and Beta: The equipoint exists and is unique.*
- (2) *Bin: Given  $\mathfrak{s}$ , if  $e_{\mathfrak{s}, \mathfrak{t}}$  is an equipoint, then  $\mathfrak{s}$  is a median for  $\text{Bin}(\mathfrak{d}, e_{\mathfrak{s}, \mathfrak{t}})$ .*
- (3) *Bin: For even  $\mathfrak{d}$  and any integer  $0 \leq k \leq \frac{\mathfrak{d}}{2}$ ,*

$$P_{\sigma(\frac{\mathfrak{d}}{2}+k)}\left(\mathfrak{S} = \frac{\mathfrak{d}}{2} + k\right) = P_{\sigma(\frac{\mathfrak{d}}{2}-k)}\left(\mathfrak{S} = \frac{\mathfrak{d}}{2} - k\right).$$

*Also we have the symmetry*

$$P_{\frac{\mathfrak{d}}{2}+k}\left(\mathfrak{S} = \frac{\mathfrak{d}}{2} + k\right) = P_{\frac{\mathfrak{d}}{2}-k}\left(\mathfrak{S} = \frac{\mathfrak{d}}{2} - k\right).$$

*Proof.* (1) Note that for fixed integer  $\mathfrak{s}$ , the function  $P_p(\mathfrak{S} \geq \mathfrak{s})$  is increasing from  $p = 0$  to  $p = 1$ . Likewise  $P_p(\mathfrak{S} \leq \mathfrak{s})$  is decreasing from  $p = 0$  to  $p = 1$ . The graphs are continuous, so must cross at a unique point, namely at  $e_{\mathfrak{s}, \mathfrak{t}}$ . Likewise,  $P^{b(\mathfrak{s}+1, \mathfrak{t})}(\mathfrak{B} \leq p)$  increases from 0 up to 1 while  $P^{b(\mathfrak{s}, \mathfrak{t}+1)}(\mathfrak{B} \geq p)$  decreases from 1 down to 0.

(2) Fix  $\mathfrak{d}, s$ , hence  $e_{\mathfrak{s}, \mathfrak{t}}$ . Then by the definition of  $e_{\mathfrak{s}, \mathfrak{t}}$ , we have

$$(15.2) \quad 1 = 2P_{e_{\mathfrak{s}, \mathfrak{t}}}(\mathfrak{S} < \mathfrak{s}) + P_{e_{\mathfrak{s}, \mathfrak{t}}}(\mathfrak{S} = \mathfrak{s}).$$

If  $P_{e_{\mathfrak{s}, \mathfrak{t}}}(\mathfrak{S} < \mathfrak{s}) + P_{e_{\mathfrak{s}, \mathfrak{t}}}(\mathfrak{S} = \mathfrak{s}) < \frac{1}{2}$  then  $P_{e_{\mathfrak{s}, \mathfrak{t}}}(\mathfrak{S} > \mathfrak{s}) + P_{e_{\mathfrak{s}, \mathfrak{t}}}(\mathfrak{S} = \mathfrak{s}) < \frac{1}{2}$  which contradicts (15.2). Thus  $\mathfrak{s}$  is a median.

(3) The symmetry is seen by switching the roles of heads and tails:

$$P_p(\mathfrak{S} = \mathfrak{s}) = P_{1-p}(\mathfrak{S} = \mathfrak{d} - \mathfrak{s}).$$

Then note  $\mathfrak{d} - (\frac{\mathfrak{d}}{2} + \frac{k}{\mathfrak{d}}) = \frac{\mathfrak{d}}{2} - \frac{k}{\mathfrak{d}}$ . ■

**15.2. Monotonicity.** For  $\mathfrak{d} \in \mathbb{R}_{>0}$  fixed recall the functions

$$(15.3) \quad \Phi(\mathfrak{s}) := P^{b(\mathfrak{s}, \mathfrak{d}-\mathfrak{s}+1)}(\mathfrak{B} \leq e_{\mathfrak{s}, \mathfrak{d}-\mathfrak{s}+1}) \quad \text{and} \quad \hat{\Phi}(\mathfrak{s}) := P^{b(\mathfrak{s}, \mathfrak{d}-\mathfrak{s}+1)}\left(\mathfrak{B} \leq \frac{\mathfrak{s}}{\mathfrak{d}}\right)$$

based on the CDF of the Beta Distribution. The proof of *one step monotonicity* of these functions claimed in Theorem 1.15 from Section 1.8 is proved below in Subsubsection 15.2.1. A similar result with the CDF replaced by the PDF is established in Subsubsection 15.2.2.

**15.2.1. Monotonicity of the CDF.**

*Proof of Theorem 1.15.* (1) The claim is that  $\Phi(\mathfrak{s}) \leq \Phi(\mathfrak{s} + 1)$  for  $\mathfrak{s}, \mathfrak{d} \in \frac{1}{2}\mathbb{N}$  and  $\frac{\mathfrak{d}}{2} \leq \mathfrak{s} < \mathfrak{d} - 1$ . Recall Lemma 12.2 which says that  $f_{s,t}(\sigma)$  defined in (12.1) when evaluated at the equipoint is

$$f_{s,t}(\sigma_{s,t}) = 2 I_{\sigma_{s,t}}\left(\frac{s}{2}, 1 + \frac{t}{2}\right) - 1$$

Using the conversion  $\mathfrak{s} = \frac{s}{2}$ , we get  $\Phi(\mathfrak{s}) = \frac{f_{s,t}(\sigma_{s,t}) + 1}{2}$ . Proposition 12.8 gives two step monotonicity of  $f_{s,t}(\sigma_{s,t})$  when  $s \geq t$  which implies  $\Phi$  is one step monotone for  $\mathfrak{s} \geq \mathfrak{t}$ .

(2) We claim that  $\hat{\Phi}(\mathfrak{s}) \leq \hat{\Phi}(\mathfrak{s} + 1)$  for  $\mathfrak{s}, \mathfrak{d} \in \mathbb{R}$  with  $\frac{\mathfrak{d}}{2} \leq \mathfrak{s} < \mathfrak{d} - 1$ .

Define  $\hat{F}$  by

$$\hat{F}(\mathfrak{d}, \mathfrak{s}) = \frac{P^{b(\mathfrak{s}, \mathfrak{t}+1)}(\mathfrak{B} \leq \frac{\mathfrak{s}}{\mathfrak{d}})}{\Gamma(\mathfrak{d} + 1)} = \frac{I_{\frac{\mathfrak{s}}{\mathfrak{d}}}(\mathfrak{s}, \mathfrak{t} + 1)}{\Gamma(\mathfrak{d} + 1)} = \frac{\int_0^{\frac{\mathfrak{s}}{\mathfrak{d}}} x^{\mathfrak{s}-1} (1-x)^{\mathfrak{t}} dx}{\Gamma(\mathfrak{t} + 1) \Gamma(\mathfrak{s})}.$$

for  $\mathfrak{s} + \mathfrak{t} = \mathfrak{d}$ . Now we show that for  $\frac{\mathfrak{d}}{2} \leq \mathfrak{s} \leq \mathfrak{d} - 1$  we have  $\hat{F}(\mathfrak{d}, \mathfrak{s} + 1) \geq \hat{F}(\mathfrak{d}, \mathfrak{s})$ , equivalently  $\frac{\hat{F}(\mathfrak{d}, \mathfrak{s} + 1)}{\hat{F}(\mathfrak{d}, \mathfrak{s})} \geq 1$ .

We start by simplifying this quotient:

$$\begin{aligned} \frac{\hat{F}(\mathfrak{d}, \mathfrak{s} + 1)}{\hat{F}(\mathfrak{d}, \mathfrak{s})} &= \frac{\int_0^{\frac{\mathfrak{s}+1}{\mathfrak{d}}} x^{\mathfrak{s}}(1-x)^{\mathfrak{t}-1} dx \Gamma(\mathfrak{t} + 1) \Gamma(\mathfrak{s})}{\Gamma(\mathfrak{t}) \Gamma(\mathfrak{s} + 1) \int_0^{\frac{\mathfrak{s}}{\mathfrak{d}}} x^{\mathfrak{s}-1}(1-x)^{\mathfrak{t}} dx} \\ &= \frac{\mathfrak{t} \int_0^{\frac{\mathfrak{s}+1}{\mathfrak{d}}} x^{\mathfrak{s}}(1-x)^{\mathfrak{t}-1} dx}{\mathfrak{s} \int_0^{\frac{\mathfrak{s}}{\mathfrak{d}}} x^{\mathfrak{s}-1}(1-x)^{\mathfrak{t}} dx}. \end{aligned}$$

Thus  $\frac{\hat{F}(\mathfrak{d}, \mathfrak{s} + 1)}{\hat{F}(\mathfrak{d}, \mathfrak{s})} \geq 1$  is equivalent to

$$(15.4) \quad \mathfrak{t} \int_0^{\frac{\mathfrak{s}+1}{\mathfrak{d}}} x^{\mathfrak{s}}(1-x)^{\mathfrak{t}-1} dx \geq \mathfrak{s} \int_0^{\frac{\mathfrak{s}}{\mathfrak{d}}} x^{\mathfrak{s}-1}(1-x)^{\mathfrak{t}} dx,$$

so it suffices to prove

$$(15.5) \quad \mathfrak{t} \int_0^{\frac{\mathfrak{s}+1}{\mathfrak{d}}} x^{\mathfrak{s}}(1-x)^{\mathfrak{t}-1} dx \geq \mathfrak{s} \int_0^{\frac{\mathfrak{s}}{\mathfrak{d}}} x^{\mathfrak{s}-1}(1-x)^{\mathfrak{t}} dx - \mathfrak{t} \int_0^{\frac{\mathfrak{s}}{\mathfrak{d}}} x^{\mathfrak{s}}(1-x)^{\mathfrak{t}-1} dx.$$

As

$$(x^{\mathfrak{s}}(1-x)^{\mathfrak{t}})' = \mathfrak{s}x^{\mathfrak{s}-1}(1-x)^{\mathfrak{t}} dx - \mathfrak{t}x^{\mathfrak{s}}(1-x)^{\mathfrak{t}-1},$$

the right-hand side of (15.5) equals

$$\frac{\mathfrak{s}^{\mathfrak{s}} \mathfrak{t}^{\mathfrak{t}}}{\mathfrak{d}^{\mathfrak{d}}}.$$

Letting  $\eta(x) := x^{\mathfrak{s}}(1-x)^{\mathfrak{t}-1}$ , we see

$$\eta'(x) = x^{\mathfrak{s}-1}(1-x)^{\mathfrak{t}-2}(-x\mathfrak{d} + x + \mathfrak{s}),$$

so  $\eta(x)$  is increasing on  $\left[0, \frac{\mathfrak{s}}{\mathfrak{d}-1}\right]$  and decreasing on  $\left[\frac{\mathfrak{s}}{\mathfrak{d}-1}, 1\right]$ . Since  $\mathfrak{s} \leq \mathfrak{d} - 1$ , we have

$\frac{\mathfrak{s}}{\mathfrak{d}-1} \in \left[\frac{\mathfrak{s}}{\mathfrak{d}}, \frac{\mathfrak{s}+1}{\mathfrak{d}}\right]$ . We claim that

$$(15.6) \quad \eta\left(\frac{\mathfrak{s}}{\mathfrak{d}}\right) \leq \eta\left(\frac{\mathfrak{s}+1}{\mathfrak{d}}\right).$$

Indeed, (15.6) is easily seen to be equivalent to

$$\left(1 + \frac{1}{\mathfrak{s}}\right)^{\mathfrak{s}} \geq \left(1 + \frac{1}{\mathfrak{t}-1}\right)^{\mathfrak{t}-1},$$

which holds since  $\mathfrak{s} \geq \mathfrak{t}$ .

We can now apply a box inequality on the left-hand side of (15.5):

$$\mathfrak{t} \int_0^{\frac{\mathfrak{s}+1}{\mathfrak{d}}} x^{\mathfrak{s}}(1-x)^{\mathfrak{t}-1} dx \geq \mathfrak{t} \frac{1}{\mathfrak{d}} \eta\left(\frac{\mathfrak{s}}{\mathfrak{d}}\right) = \frac{\mathfrak{s}^{\mathfrak{s}} \mathfrak{t}^{\mathfrak{t}}}{\mathfrak{d}^{\mathfrak{d}}},$$

establishing (15.5). ■

Ideas in the paper [PR07] were very helpful in the proof above.

15.2.2. *Monotonicity of the PDF.* So far we have studied the CDF of the Beta Distribution. However, the functions

$$(15.7) \quad P_{e_{\mathfrak{s},\mathfrak{t}}}(\mathfrak{S} = \mathfrak{s}) \quad \text{and} \quad P_{\frac{\mathfrak{s}}{\mathfrak{d}}}(\mathfrak{S} = \mathfrak{s})$$

based on PDF's of the Binomial distribution also have monotonicity properties for integer  $\mathfrak{d}/2 \leq \mathfrak{s} \leq \mathfrak{d}$ .

**Proposition 15.2.** *Let  $\mathfrak{d} \in \mathbb{N}$ . For integer  $\mathfrak{s} \geq \frac{\mathfrak{d}}{2}$ , we have that*

- (1)  $P_{\frac{\mathfrak{s}}{\mathfrak{d}}}(\mathfrak{S} = \mathfrak{s})$  is increasing; its minimum is  $P_{\frac{\mathfrak{s}}{\mathfrak{d}}}(\mathfrak{S} = \lceil \mathfrak{d}/2 \rceil)$ ;
- (2)  $P_{e_{\mathfrak{s},\mathfrak{t}}}(\mathfrak{S} = \mathfrak{s})$  is increasing; its minimum is  $P_{\sigma_{\mathfrak{s}}}(\mathfrak{S} = \lceil \mathfrak{d}/2 \rceil)$ .

*Proof.* (2) By the definition of  $e_{\mathfrak{s},\mathfrak{t}}$ , we have  $P_{e_{\mathfrak{s},\mathfrak{t}}}(\mathfrak{S} = \mathfrak{s}) = 2P_{e_{\mathfrak{s},\mathfrak{t}}}(\mathfrak{S} \leq \mathfrak{s}) - 1$ . Theorem 1.15 implies the required monotonicity.

(1) Recall

$$P_{\frac{\mathfrak{s}}{\mathfrak{d}}}(\mathfrak{S} = \mathfrak{s}) = \binom{\mathfrak{d}}{\mathfrak{s}} \frac{\mathfrak{s}^{\mathfrak{s}} \mathfrak{t}^{\mathfrak{t}}}{\mathfrak{d}^{\mathfrak{d}}}.$$

Thus

$$\frac{P_{\frac{\mathfrak{s}+1}{\mathfrak{d}}}(\mathfrak{S} = \mathfrak{s} + 1)}{P_{\frac{\mathfrak{s}}{\mathfrak{d}}}(\mathfrak{S} = \mathfrak{s})} = \left( \frac{\mathfrak{s} + 1}{\mathfrak{s}} \right)^{\mathfrak{s}} \left( \frac{\mathfrak{t} - 1}{\mathfrak{t}} \right)^{\mathfrak{t}-1}$$

is  $\geq 1$  iff

$$(15.8) \quad \left( 1 + \frac{1}{\mathfrak{s}} \right)^{\mathfrak{s}} \geq \left( 1 + \frac{1}{\mathfrak{t} - 1} \right)^{\mathfrak{t}-1}.$$

Since  $\mathfrak{s} > \mathfrak{t} - 1$ , (15.8) holds, establishing the monotonicity of  $P_{\frac{\mathfrak{s}}{\mathfrak{d}}}(\mathfrak{S} = \mathfrak{s})$ . ■

## REFERENCES

- [Arv69] W.B. Arveson: Subalgebras of  $C^*$ -algebras, *Acta Math.* **123** (1969) 141–224. [2](#), [3](#), [8](#)
- [Arv72] W.B. Arveson: Subalgebras of  $C^*$ -algebras II, *Acta Math.* **128** (1972) 271–308. [2](#), [3](#), [8](#)
- [Arv08] W.B. Arveson: The noncommutative Choquet boundary, *J. Amer. Math. Soc.* **21** (2008) 1065–1084. [6](#)
- [BB13] C. Badea, B. Beckermann: Spectral Sets, In: Second edition of *Handbook of Linear Algebra* (L. Hogben, ed.). CRC Press, 2013. [3](#)
- [Bal11] J.A. Ball: Multidimensional circuit synthesis and multivariable dilation theory, *Multidimens. Syst. Signal Process.* **22** (2011) 27–44. [2](#), [3](#)
- [BGR90] J.A. Ball, I. Gohberg, L. Rodman: *Interpolation of rational matrix functions*, Operator Theory: Advances and Applications **45**, Birkhäuser Verlag, 1990. [2](#), [3](#)
- [BGKP+] A. Belton, D. Guillot, A. Khare, M. Putinar: Schoenberg’s positivity theorem in fixed dimension, preprint <http://arxiv.org/abs/1504.07674> [9](#)
- [B-TN02] A. Ben-Tal, A. Nemirovski: On tractable approximations of uncertain linear matrix inequalities affected by interval uncertainty, *SIAM J. Optim.* **12** (2002) 811–833. [2](#), [4](#), [9](#), [16](#), [42](#), [72](#)
- [BLM04] D.P. Blecher, C. Le Merdy: *Operator algebras and their modules – an operator space approach*, London Mathematical Society Monographs **30**, Oxford University Press, 2004. [2](#)
- [BPR13] G. Blekherman, P.A. Parrilo, R.R. Thomas (editors): *Semidefinite optimization and convex algebraic geometry*, MOS-SIAM Series on Optimization **13**, SIAM, 2013. [2](#), [6](#)
- [BGFB94] S. Boyd, L. El Ghaoui, E. Feron, V. Balakrishnan: *Linear Matrix Inequalities in System and Control Theory*, SIAM Studies in Applied Mathematics **15**, SIAM, 1994. [2](#), [6](#)
- [Der06] J. Dereziński: Introduction to Representations of the Canonical Commutation and Anticommutation Relations, In: *Large Coulomb Systems*, 63–143, Springer, 2006. [75](#)
- [DZ91] P. Diaconis, S. Zabel: Closed form summation for classical distributions: variations on a theme of de Moivre, *Statistical Science* **6** (1991) 284–302. [59](#)
- [Dav12] K.R. Davidson: The mathematical legacy of William Arveson. *J. Operator Theory* **68** (2012) 307–334. [2](#)
- [DDSS+] K. R. Davidson, A. Dor-On, O. Shalit, B. Solel: Dilations, inclusions of matrix convex sets, and completely positive maps, preprint <http://arxiv.org/abs/1601.07993> [15](#)
- [DK+] K.R. Davidson, M. Kennedy: The Choquet boundary of an operator system, to appear in *Duke Math. J.* **164** (2015) 2989–3004. [2](#)
- [dOHMP09] M. de Oliveira, J.W. Helton, S. McCullough, M. Putinar: Engineering systems and free semi-algebraic geometry, in: *Emerging applications of algebraic geometry* (edited by M. Putinar, S. Sullivant), 17–61, Springer-Verlag, 2009. [2](#)
- [EW97] E.G. Effros, S. Winkler: Matrix convexity: operator analogues of the bipolar and Hahn-Banach theorems, *J. Funct. Anal.* **144** (1997) 117–152. [2](#), [40](#), [41](#)
- [FP12] D. Farenick, V.I. Paulsen: Operator system quotients of matrix algebras and their tensor products, *Math. Scand.* **111** (2012) 210–243. [6](#)
- [FFGK98] C. Foias, A.E. Frazho, I. Gohberg, M.A. Kaashoek: *Metric constrained interpolation, commutant lifting and systems*, Operator Theory: Advances and Applications **100**, Birkhäuser Verlag, 1998. [3](#)
- [GäWg54] L. Gårding, A. Wightman: Representations of the anticommutation relations, *Proc. Nat. Acad. Sci.* **40** (1954) 617–621. [75](#)
- [GoW195] M.X. Goemans, D.P. Williamson: Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming, *Journal of the ACM* **42** (1995) 1115–1145. [9](#)
- [GM77] R.A. Groeneveld, G. Meeden: The mode, median, and mean inequality, *The American Statistician* **31** (1977) 120–121. [13](#)
- [HKM12] J.W. Helton, I. Klep, S. McCullough: The convex Positivstellensatz in a free algebra, *Adv. Math.* **231** (2012) 516–534. (this article succeeds [HKM13] but appeared earlier) [3](#), [7](#), [87](#)
- [HKM13] J.W. Helton, I. Klep, S. McCullough: The matricial relaxation of a linear matrix inequality, *Math. Program.* **138** (2013) 401–445. (this article precedes [HKM12] but appeared later) [3](#), [7](#), [8](#), [9](#), [79](#), [80](#), [87](#)

- [HKM+] J.W. Helton, I. Klep, S. McCullough: The Tracial Hahn-Banach Theorem, Polar Duals, Matrix Convex Sets, and Projections of Free Spectrahedra, to appear in *J. Eur. Math. Soc.* <http://arxiv.org/abs/1407.8198>. 80
- [HLLL14a] D. Han, D.R. Larson, B. Liu, R. Liu: *Operator-valued measures, dilations, and the theory of frames*, Memo. Amer. Math. Soc. **1075** (2014) 1–98. 3
- [HLLL14b] D. Han, D.R. Larson, B. Liu, R. Liu: Dilations of frames, operator-valued measures and bounded linear maps, In: *Operator methods in wavelets, tilings, and frames*, 33–53, Contemp. Math. **626**, Amer. Math. Soc., 2014. 3
- [HM12] J.W. Helton, S. McCullough: Every free basic convex semi-algebraic set has an LMI representation, *Ann. of Math. (2)* **176** (2012) 979–1013. 2, 41
- [HV07] J.W. Helton, V. Vinnikov: Linear matrix inequality representation of sets, *Comm. Pure Appl. Math.* **60** (2007) 654–674. 6
- [KVV14] D. Kaliuzhnyi-Verbovetskyi, V. Vinnikov: *Foundations of Free Noncommutative Function Theory*, Mathematical Surveys and Monographs **199**, AMS, 2014. 2
- [KPTT13] A.S. Kavruk, V.I. Paulsen, I.G. Todorov, M. Tomforde: Quotients, exactness, and nuclearity in the operator system category, *Adv. Math.* **235** (2013) 321–360. 6
- [KV71] J.D. Kečkić, P.M. Vasić: Some inequalities for the gamma function, *Publ. Inst. Math., Belgrade* **25** (1971) 107–114. 49
- [KTT13] K. Kellner, T. Theobald, C. Trabant: Containment problems for polytopes and spectrahedra, *SIAM J. Optim.* **23** (2013) 1000–1020. 9
- [Ker+] J. Kerman: A closed-form approximation for the median of the beta distribution, *preprint* <http://arxiv.org/abs/1111.0433v1> 57
- [LS13] D.R. Larson and F.H. Szafraniec: Framings and dilations, *Acta Sci. Math (Szeged)* **79** (2013) 529–543. 3
- [MV70] D.S. Mitrinović, P.M. Vasić: *Analytic inequalities*, Springer-Verlag, 1970. 65
- [Nai43] M.A. Naimark: Positive definite operators on a commutative group, *Izv. Akad. Nauk SSSR Ser. Mat.* **7** (1943) 237–244. 3
- [Nes97] Yu. Nesterov: Semidefinite relaxation and nonconvex quadratic optimization, *Optim. Methods Softw.* **9** (1998) 141–160. 2, 9
- [Nem06] A. Nemirovskii: Advances in convex optimization: conic programming, *plenary lecture*, International Congress of Mathematicians (ICM), Madrid, Spain, 2006. 2, 4, 9
- [NC11] M.A. Nielsen, I.L. Chuang: *Quantum Computation and Quantum Information*, Cambridge Univ. Press, 2011. 3
- [Pau02] V. Paulsen: *Completely bounded maps and operator algebras*, Cambridge Univ. Press, 2002. 2, 3, 8, 9, 10, 42, 72, 75
- [PYY89] M.E. Payton, L.J. Young, J.H. Young: Bounds for the difference between median and mean of beta and negative binomial distributions, *Metrika* **36** (1989) 347–354. 13, 57, 61
- [PR07] O. Perrin, E. Redside: Generalization of Simmons’ Theorem, *Statist. Probab. Lett.* **77** (2007) 604–606. 3, 13, 14, 44, 86
- [Pis03] G. Pisier: *Introduction to operator space theory*, London Mathematical Society Lecture Note Series **294**, Cambridge Univ. Press, 2003. 2, 10, 72, 73, 75
- [Rai71] E.D. Rainville: *Special functions*, Chelsea Publishing Comp., 1971. 29
- [Tay72] J.L. Taylor: A general framework for a multi-operator functional calculus, *Adv. Math.* **9** (1972) 183–252. 2
- [Tay73] J.L. Taylor: Functions of several noncommuting variables, *Bull. Amer. Math. Soc.* **79** (1973) 1–34. 2
- [SIG96] R.E. Skelton, T. Iwasaki, K.M. Grigoriadis: *A Unified Algebraic Approach to Linear Control Design*, Taylor and Francis, 1996. 2
- [Sim1894] T.C. Simmons: A New Theorem in Probability, *Proc. London Math. Soc.* **1** (1894) 290–325. 3, 6
- [Sti55] W.F. Stinespring: Positive functions on  $C^*$ -algebras, *Proc. Amer. Math. Soc.* **6** (1955) 211–216. 8
- [SzN53] B. Sz-Nagy: Sur contractions le space Hilbert, *Acta Sci. Math. (Szeged)* **15** (1953) 87–92. 3



- [SzNFBK10] B. Sz-Nagy, C. Foias, H. Bercovici, L. Kerchy: *Harmonic Analysis of Operators on Hilbert Space*, Springer-Verlag, 2010. [3](#)
- [Voi04] D.-V. Voiculescu: Free analysis questions I: Duality transform for the coalgebra of  $\partial_{X:B}$ , *Int. Math. Res. Not.* **16** (2004) 793–822. [2](#)
- [Voi10] D.-V. Voiculescu: Free analysis questions II: The Grassmannian completion and the series expansions at the origin, *J. reine angew. Math.* **645** (2010) 155–236. [2](#)
- [WSV00] H. Wolkowicz, R. Saigal, L. Vandenberghe (editors): *Handbook of semidefinite programming. Theory, algorithms, and applications*, Kluwer Academic Publishers, 2000. [2](#), [4](#), [7](#)
- [Zal+] A. Zalar: Operator Positivstellensätze for noncommutative polynomials positive on matrix convex sets, *preprint* <http://arXiv.org/abs/1602.00765>. [15](#)

J. WILLIAM HELTON, DEPARTMENT OF MATHEMATICS, UNIVERSITY OF CALIFORNIA, SAN DIEGO  
*E-mail address:* [helton@math.ucsd.edu](mailto:helton@math.ucsd.edu)

IGOR KLEP, DEPARTMENT OF MATHEMATICS, THE UNIVERSITY OF AUCKLAND, NEW ZEALAND  
*E-mail address:* [igor.klep@auckland.ac.nz](mailto:igor.klep@auckland.ac.nz)

SCOTT MCCULLOUGH, DEPARTMENT OF MATHEMATICS, UNIVERSITY OF FLORIDA, GAINESVILLE  
*E-mail address:* [sam@math.ufl.edu](mailto:sam@math.ufl.edu)

MARKUS SCHWEIGHOFER, FACHBEREICH MATHEMATIK UND STATISTIK, UNIVERSITÄT KONSTANZ  
*E-mail address:* [markus.schweighofer@uni-konstanz.de](mailto:markus.schweighofer@uni-konstanz.de)

## CONTENTS

1. Introduction .....	2
1.1. Simultaneous dilations .....	3
1.2. Solution of the minimization problem (1.1) .....	4
1.2.1. Proof of Theorem 1.2 .....	5
1.2.2. Coin flipping and Simmons' Theorem .....	5
1.3. Linear matrix inequalities (LMIs), spectrahedra and general dilations .....	6
1.3.1. Dilations to commuting operators .....	7
1.3.2. Spectrahedral inclusion problem .....	7
1.3.3. Accuracy of the free relaxation .....	7
1.4. Interpretation in terms of completely positive maps .....	8
1.5. Matrix cube problem .....	9
1.6. Matrix balls .....	9
1.7. Adapting the Theory to Free Nonsymmetric Variables .....	10
1.8. Probabilistic theorems and interpretations .....	11
1.8.1. Binomial distributions .....	11
1.8.2. Equipoints and medians .....	12
1.8.3. Equipoints compared to medians .....	13
1.8.4. Monotonicity of the CDF .....	14
1.9. Reader's guide .....	14
2. Dilations and Free Spectrahedral Inclusions .....	15
3. Lifting and Averaging .....	17
4. A Simplified Form for $\vartheta$ .....	19
5. $\vartheta$ is the Optimal Bound .....	23
5.1. Averages over $O(d)$ equal averages over $S^{d-1}$ .....	23
5.2. Properties of matrices gotten as averages .....	24
5.3. Dilating to commuting self-adjoint operators .....	25
5.4. Optimality of $\kappa_*(d)$ .....	27
6. The Optimality Condition $\alpha = \beta$ in Terms of Beta Functions .....	29
7. Rank versus Size for the Matrix Cube .....	33
7.1. Proof of Theorem 1.6 .....	36
8. Free Spectrahedral Inclusion Generalities .....	37
8.1. A general bound on the inclusion scale .....	38
8.2. The inclusion scale equals the commutability index .....	40
8.2.1. Matricial Hahn-Banach background .....	40
8.2.2. Proof of Theorem 8.4 .....	41

8.2.3. Matrix cube revisited .....	42
9. Reformulation of the Optimization Problem.....	43
10. Simmons' Theorem for Half Integers.....	44
10.1. Two step monotonicity of $c_s$ .....	45
10.2. The upper boundary case.....	48
10.3. The lower boundary cases for $d$ even.....	50
10.4. The lower boundary cases for $d$ odd .....	51
11. Bounds on the Median and the Equipoint of the Beta Distribution.....	56
11.1. Lower bound for the equipoint $e_{s,t}$ .....	57
11.2. New bounds on the median of the beta distribution.....	58
12. Proof of Theorem 1.2.....	62
12.1. An auxiliary function.....	62
12.2. Two step monotonicity of $f_{s,t}(\sigma_{s,t})$ .....	63
12.3. Boundary cases.....	65
13. Estimating $\vartheta(d)$ for Odd $d$ . .....	67
13.1. Proof of Theorem 13.1.....	68
13.2. Explicit bounds on $\vartheta(d)$ .....	71
14. Dilations and Inclusions of Balls .....	72
14.1. The general dilation result.....	72
14.2. Four types of balls.....	73
14.2.1. The OH ball.....	73
14.2.2. The min and max balls .....	74
14.2.3. The spin ball and the canonical anticommutation relations.....	75
14.3. Inclusions and dilations.....	78
14.3.1. An alternate proof of Proposition 14.14.....	82
15. Probabilistic Theorems and Interpretations continued.....	83
15.1. The nature of equipoints.....	83
15.2. Monotonicity .....	84
15.2.1. Monotonicity of the CDF .....	84
15.2.2. Monotonicity of the PDF .....	86
References .....	87
Index .....	92

$C_D$ , 18  
 $E_J$ , 23  
 $J(s, t; a, b)$ , 19  
 $J_*$ , 25  
 $V$ , 17  
 $\alpha(s, t; a, b) := \frac{1}{d} \left( 2I_{\frac{a}{a+b}} \left( \frac{t}{2}, \frac{s}{2} + 1 \right) - 1 \right)$ , 31  
 $\alpha$ , 20  
 $\alpha(s, t; a, b)$ , 20  
 $\beta(s, t; a, b) := \frac{1}{d} \left( 2I_{\frac{b}{a+b}} \left( \frac{s}{2}, \frac{t}{2} + 1 \right) - 1 \right)$ , 31  
 $\beta$ , 20  
 $\beta(s, t; a, b)$ , 20  
 $\mathcal{D}_L$ , 2  
 $\mathfrak{C}^{(g)}$ , 2  
 $\kappa_*(s, t)$ , 20  
 $\kappa(s, t; a, b)$ , 20  
 $\mathbb{B}_g$ , 73  
 $\mathcal{D}_L$ - $\mathcal{D}_{\bar{L}}$ -inclusion constant, 7  
 $\mathcal{D}_L$ -inclusion scale, 7  
 $\mathcal{H}$ , 17  
 $\mathfrak{B}_g^{\max}$ , 74  
 $\mathfrak{B}_g^{\min}$ , 74  
 $\mathfrak{B}_g^{\text{oh}}$ , 73  
 $\mathfrak{B}_g^{\text{spin}}$ , 76  
 $\mathcal{S}_L$ , 2  
 $\mu_{s,t} := \frac{s}{s+t}$ , 57  
 $\rho_g(d)$ , 42  
 $\text{sign}_0(B) = (s, t)$ , 19  
 $\tau_g(d)$ , 42  
 $\vartheta(d)$ , 2  
  
 averaging, 17, 23, 24  
  
 beta distribution, 11, 56, 58, 83  
 beta function, 4  
 beta function, incomplete, 4  
 beta function, regularized, 4  
 binomial distribution, 83  
  
 canonical anticommutation relations (CAR), 75  
 closed under direct sums, 40  
 closed under simultaneous conjugation by  
     contractions, 40  
 coin flipping, 11, 12  
 commutability index, 7, 40  
 completely positive map, 3  
 compression, 3  
 contraction, 4  
 cube, 6, 16, 27, 42

## INDEX

dilation, 2, 3, 7, 15, 17, 27  
     Halmos, 82  
     Naimark, 3  
 distribution, 11  
     Beta, 11  
     Binomial, 11  
 dual  
     polar, 75  
  
 equipoint, 12, 13, 44, 56, 83  
 Euler function, 29  
 Euler gamma function, 29  
 extreme point, 80  
  
 free analysis, 2  
 free cube, 2, 6, 16, 27, 42  
 free relaxation, 9  
 free spectrahedral inclusion, 15, 37  
 free spectrahedral inclusion problem, 7  
 free spectrahedron, 6  
  
 gamma function, 29  
  
 inclusion scale, 38, 40  
 incomplete beta function, 4  
  
 linear matrix inequality, 2, 6  
 linear pencil, 2, 6  
 LMI, 2, 6  
 LMI domain, 6  
  
 matricial relaxation, 9  
 matrix ball, 9  
 matrix ball problem, 80  
 matrix convex, 40  
 matrix cube problem, 9  
 max ball, 74  
 median, 12, 13, 56, 58, 83  
 min ball, 74  
 monic linear pencil, 6  
  
 OH ball, 73  
  
 polar dual, 75  
 polyhedron, 6  
  
 real Reinhardt, 73  
 regularized beta function, 4  
  
 SDP, 7

- semidefinite programming, [2](#), [7](#)
- signature, [26](#)
- signature matrix, [25](#)
- spectrahedron, [2](#), [6](#)
- spin ball, [76](#)
- spin system, [75](#)
- symmetry matrix, [26](#)
  
- tensor product, [6](#)
  - Kronecker, [6](#)
- theorem
  - Ben-Tal, Nemirovski, [9](#)
  - Effros-Winkler, [40](#)
  - Goemans-Williamson, [9](#)
  - Grothendieck, [9](#)
  - Hahn-Banach, [40](#)
  - Nesterov's  $\frac{\pi}{2}$ , [9](#)
  - separation of matrix convex sets, [40](#)
  - Simmons', [13](#), [44](#)
  - simultaneous dilation, [10](#)
  - Sz.-Nagy, [3](#)
  - von Neumann, [3](#)
- von Neumann's inequality, [3](#)